8-2011

CLASSIFICATION FOR MASS SPECTRA AND COMPREHENSIVE TWO-DIMENSIONAL CHROMATOGRAMS

Xue Tian
University of Nebraska - Lincoln, xtian@cse.unl.edu

CLASSIFICATION FOR MASS SPECTRA AND COMPREHENSIVE TWO-DIMENSIONAL CHROMATOGRAMS

by

Xue Tian

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Computer Science

Under the Supervision of Professor Stephen E. Reichenbach

Lincoln, Nebraska

August, 2011

CLASSIFICATION FOR MASS SPECTRA AND COMPREHENSIVE TWO-DIMENSIONAL CHROMATOGRAMS

Xue Tian, Ph.D.

University of Nebraska, 2011

Adviser: Stephen E. Reichenbach

Mass spectra contain characteristic information regarding the molecular structure and properties of compounds. The mass spectra of compounds from the same chemically related group are similar. Classification is one of the fundamental methodologies for analyzing mass spectral data. The primary goals of classification are to automatically group compounds based on their mass spectra, to find correlation between the properties of compounds and their mass spectra, and to provide a positive identification of unknown compounds.

This dissertation presents a new algorithm for the classification of mass spectra, the most similar neighbor with a probability-based spectrum similarity measure (MSN-PSSM). Experimental results demonstrate the effectiveness and robustness of the new MSN-PSSM algorithm. In leave-one-out cross-validation, it outperforms popular techniques for classification of mass spectra, such as principal component analysis with discriminant function analysis, soft independent modeling of class analogy, and decision tree learning.

Comprehensive two-dimensional chromatography yields highly informative separation patterns because of its great practical peak capacity and sensitivity produced by applying two different separation principles. However, the improvement in information yields complex data requiring comprehensive analyses to interpret the rich information and to extract

useful information for characterizing sample composition.

This dissertation presents a new non-targeted cross-sample classification method to analyze comprehensive two-dimensional chromatograms. Experimental results validate the effectiveness of the new non-targeted cross-sample classification. The new non-targeted cross-sample classification is successfully applied to a set of comprehensive two-dimensional chromatograms of breast cancer tumor samples. The feature vectors generated by the new non-targeted cross-sample classification are useful for discriminating between breast cancer tumor samples of different grades and providing information to identify potential biomarkers for closer examination.

Keywords: classification, mass spectrometry, comprehensive two-dimensional chromatography.

ACKNOWLEDGMENTS

First of all, I would like to express my sincere thanks to my advisor Dr. Stephen E. Reichenbach for his inspiring and encouraging guidance at every aspect of my academic work and research. The work presented in this dissertation would not be possible without his inspirational ideas, invaluable insights, persistent encouragement, and overwhelming enthusiasm. I also thank him for his patience that helped me overcome many difficulties and guided me toward achieving my research goals.

I would like to thank my doctoral committee members Dr. Berthe Y. Choueiry, Dr. Ashok K. Samal, Dr. Stephen D. Scott, and Dr. Hendrik J. Viljoen who have been giving me insightful advice and vital assistance on my course work and research. I also thank them for their constructive comments on my proposal and dissertation.

A special thank to Dr. Qingping Tao from GC Image LLC for generous sharing of his knowledge and many stimulating discussions.

Thank all my colleagues and friends in the Department of Computer Science and Engineering for their help and friendship. Thank all my friends in Lincoln for all the good time we had together.

Finally, I would like to thank my husband, my parents, and my brother for their support, guidance and belief. I dedicate my doctoral dissertation to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Classification is fundamental in science uncovering inherent orders, regularities, underlying structure, and natural laws [1]. The development of new analysis and measurement techniques with high quantification ability (e.g., separation capacity) potentially improves the accuracy and precision of classification. However, these techniques often generate information-rich observations that make new mathematical or statistical methods increasingly important both for informatics and classification. Comprehensive two-dimensional gas chromatography (GC$\times$GC) with mass spectrometry (MS) and time-of-flight secondary ion mass spectrometry (TOFMS) are analysis techniques generating information-rich data. An especially important analytical task is to establish new methodologies for comprehensive information analysis and classification of these information-rich data.

1.1 Mass Spectrometry

Mass spectrometry is an analysis technique that measures the mass-to-charge ratio (m/z) of molecular and fragmentary ions [2]. The basic principle of mass spectrometry is to generate ions from compounds, to separate these ions by their m/z, and to measure them qualitatively and quantitatively by their respective m/z and intensity [3].

1.1.1 Basic Procedure of Mass Spectrometry

There are three essential steps in mass spectrometry:

1. An ionization source (commonly a beam of 70 volt electrons) converts the analyte molecules into molecular ions (M + $e \longrightarrow$ M^+ + 2e). The excess energy transferred from the ionization source leads to fragmentation to additional smaller ions.

2. A mass analyzer separates molecular ions and their charged fragments according to their m/z.

3. A detector measures the intensity of ion currents due to mass-separated ions and provides a mass spectrum.

Figure 1.1 illustrates the basic procedure of mass spectrometry.

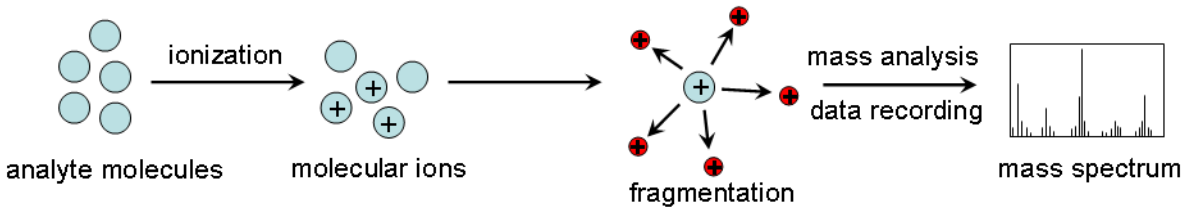

Figure 1.1: Basic procedure of mass spectrometry [2].

1.1.2 Types of Mass Spectrometers

There are many types of mass spectrometers, of which the quadrupole and the time-of-flight are the most widely used types.

1. Quadrupole employs a combination of direct-current and radio-frequency potentials as a mass "filter". As the name implies, quadrupole consists of four parallel rods arranged symmetrically to produce hyperbolic fields. Opposite rods are connected together electrically and to radio-frequency and direct-current voltage generators. Ions travel down the quadrupole region between the four rods. Only ions of a certain m/z will reach the detector for a given ratio of voltages. This mechanism allows scanning a range of m/z values by continuously varying the voltages [4].

2. Time-of-flight uses an electric field to accelerate the ions through the same potential and measures the times they take to reach the detector. All the ions receive the same kinetic energy during acceleration, but because they have different masses, they separate into groups according to velocity (and hence mass). The m/z of an ion is reflected by its time of arrival at the detector. With the same charge, ions of low mass reach the detector before those of high mass [4].

1.1.3 Nature of Mass Spectra

A mass spectrum is an array of pairs of the form (m/z, intensity). The value pairs typically are listed in ascending order from smallest to largest m/z. In an integer mass spectrum, the m/z values are integers and the intensity values are integrated to integer m/z. In a high-resolution mass spectrum, the m/z values are floating point values. A mass spectrum can be represented as a bar graph, in which each bar represents ions having a specific m/z and the length of the bar indicates the signal intensity of the ions. The most intense ion is referred to as the base peak. Typically, the ions formed in a mass spectrometer have a single charge, so the m/z value is equivalent to mass itself [3]. To compare different mass spectra, intensity values of mass spectra are normalized. Base-peak normalization is a common normalization method. In base-peak normalization, the intensity values of mass spectra are normalized to the intensity of the most intense peak (base peak) and multiplied by 999, then rounded to the closest integer value. After base-peak normalization, the intensity of the base peak is 999. Often, the largest-mass ion in a spectrum is the molecular ion (when the molecular ion is present), and smaller-mass ions are fragments from the molecular ion.

For example, the base-peak normalized mass spectrum of methanol and major ions are shown in Figure 1.2. CH_3OH^+ (the molecular ion, m/z=32) and fragment ions appear in this mass spectrum. The x-axis of this bar graph is the m/z. The y-axis is the base-peak normalized intensity of each ion, which is related to the normalized number of times an ion of that m/z strikes the detector.

Analyses often are performed on mass spectra far more complex than methanol. Interpretation of complex mass spectra is difficult or even impossible as initial fragments undergo further fragmentation and rearrangements occur.

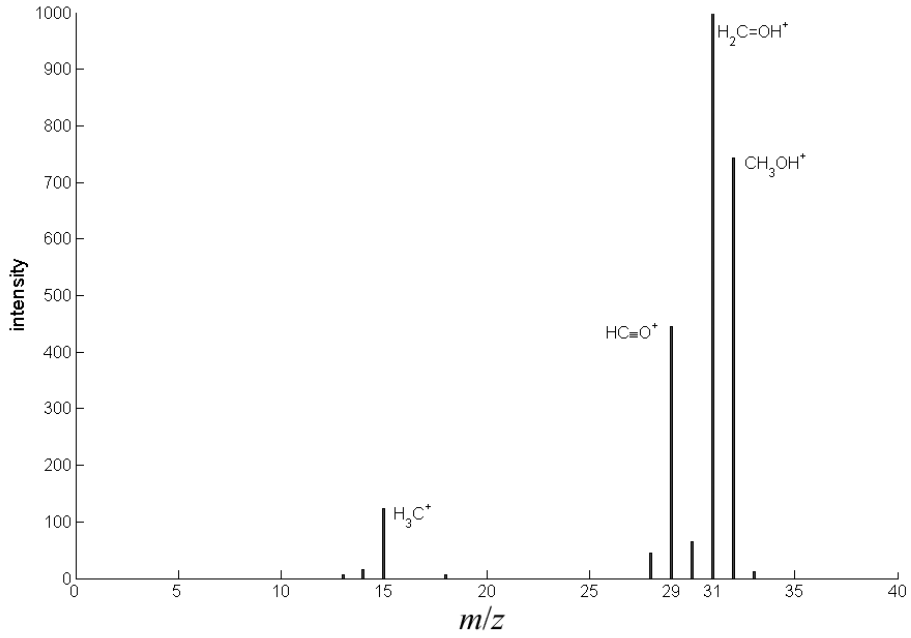

Figure 1.2: An integer mass spectrum of methanol represented as a bar graph. The x-axis is the m/z. The y-axis is the base-peak normalized intensity of each ion, which is related to the normalized number of times an ion of that m/z strikes the detector. CH_3OH^+ is the molecular ion. $H_2C{=}OH^+$, $HC{\equiv}O^+$, and H_3C^+ are fragment ions.

1.2 Comprehensive Two-Dimensional Gas Chromatography with Mass Spectrometry

Comprehensive two-dimensional gas chromatography (GC×GC) separates chemical compounds with two capillary columns coupled by a modulator as illustrated in Figure 1.3. The modulator collects and periodically injects the first column eluent (partially resolved compounds from the first column carrying out the initial separation) into a second column of different selectivity, allowing further separation [5]. Often, the first column is volatility selective and the second column is polarity selective. GC×GC provides a two-dimensional chemical ordering (by retention times) that is useful for separation of compounds in com-

plex samples and recognizing individual chemical compounds [6].

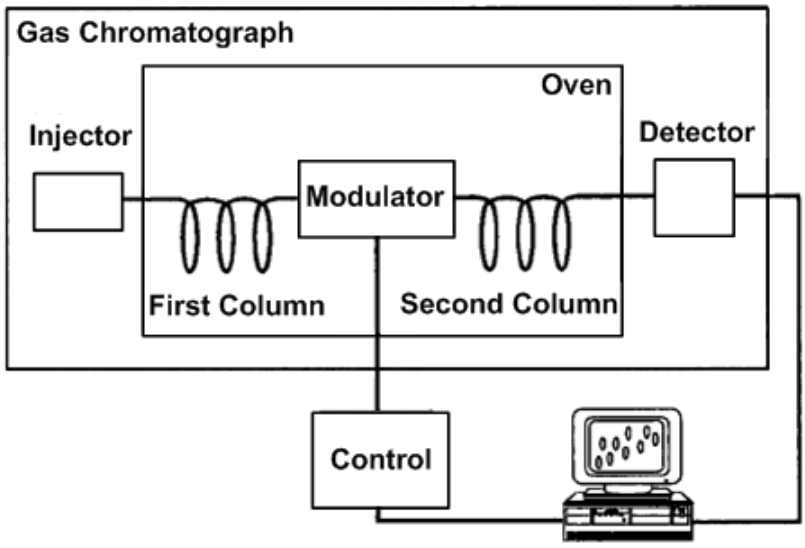

Figure 1.3: GC×GC system separates chemical compounds with two capillary columns coupled by a modulator.

The eluent of the second column can be input to a mass spectrometer to produce rich structural information for chemical identification. GC×GC with mass spectrometry (GC×GC-MS) provides large separation capability (capable of resolving several thousands of chemical compounds) and large capability for identifying chemical constituents of highly complex mixtures.

GC×GC-MS output is a three-way data cube (each way is functionally linked, that is, the output from one way modulates the output of subsequent ways) [7]. The first way is the elapsed time for the first column separation; the second way is the elapsed time for the second column separation; and the third way is the mass spectrum.

GC×GC-MS data can be displayed as a two-dimensional total ion count (TIC) image. In the TIC image, the x-axis is the elapsed time for the first column separation; the y-axis

is the elapsed time for the second column separation. The value of each pixel of the TIC image is the summation of all the intensity values of the mass spectrum associated with the pixel. Figure 1.4 illustrates the GC×GC-MS TIC image of a mixture of compounds containing paraffins, isoparaffins, aromatics, naphthenes, and olefins (PIANO) obtained from Supelco, Inc. Each compound produces a two-dimensional peak (represented as a group of pixels) of adjacent pixels with larger pixel values than the surrounding pixels in the TIC image.

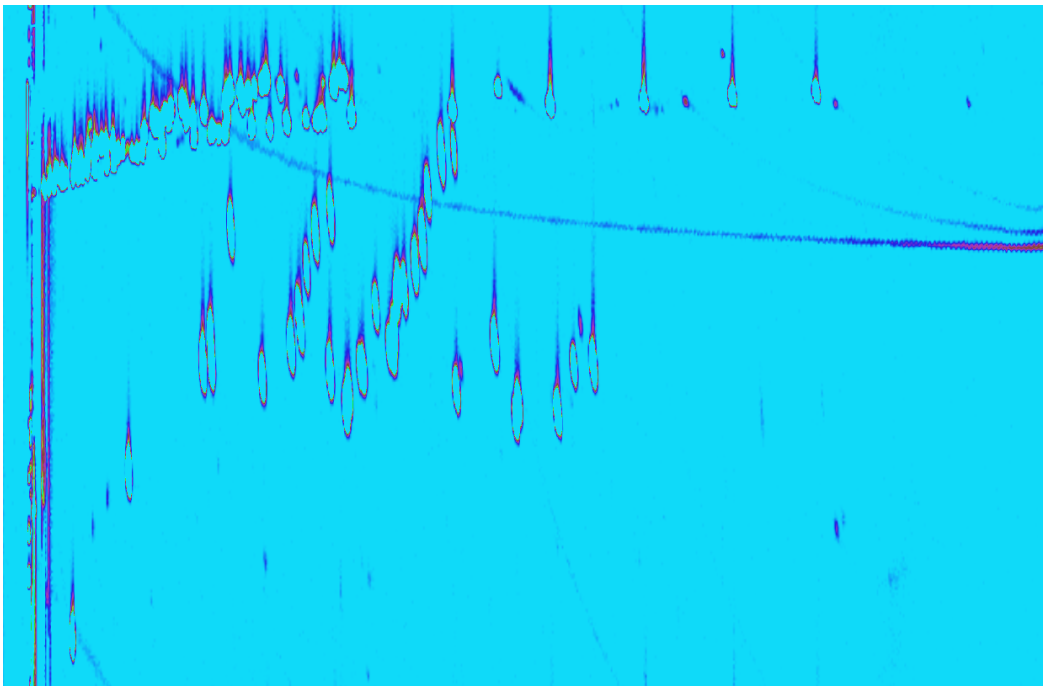

Figure 1.4: GC×GC-MS TIC image of a PIANO mixture. The x-axis is the elapsed time for the first column separation; the y-axis is the elapsed time for the second column separation. The value of each pixel of the TIC image is the summation of all the intensity values of the mass spectrum associated with the pixel.

1.3 Time-of-Flight Secondary Ion Mass Spectrometry

Time-of-flight secondary ion mass spectrometry (ToF-SIMS) is an analysis technique characterizing the surface and near surface ($\sim$30μm) region of solids and the surface of some liquids [8]. In ToF-SIMS, the target is placed in an ultra-high vacuum environment where a beam of energetic (between 0.5 keV and 20 keV) primary ions (e.g., Au_3^+ or C_{60}^+) bombards the target surface and sputters atoms, molecules, and molecular fragments (secondary ions) from the target surface as Figure 1.5 [9] illustrates. The sputtering process consists of the implantation of the primary ions into the target and the removal of surface atoms by the energy loss of the primary ions in the form of a collision cascade [8]. The secondary ions ejected from the target surface are electrostatically accelerated into a "flight tube" and separated according to their m/z which is determined by measuring the times at which they reach the detector (time-of-flight) [10].

The primary ion beam can be finely focused to sweep the target surface in a raster pattern at a submicrometer lateral resolution to create a two-dimensional ToF-SIMS TIC image. In the ToF-SIMS TIC image, the value of each pixel is the summation of all the intensity values of the mass spectrum associated with the pixel. Repeating the raster pattern to drill into the target can create a three-dimensional ToF-SIMS TIC image. Figure 1.6 illustrates the three-dimensional ToF-SIMS TIC image of a frog oocyte obtained from the Surface Analysis Research Centre, University of Manchester.

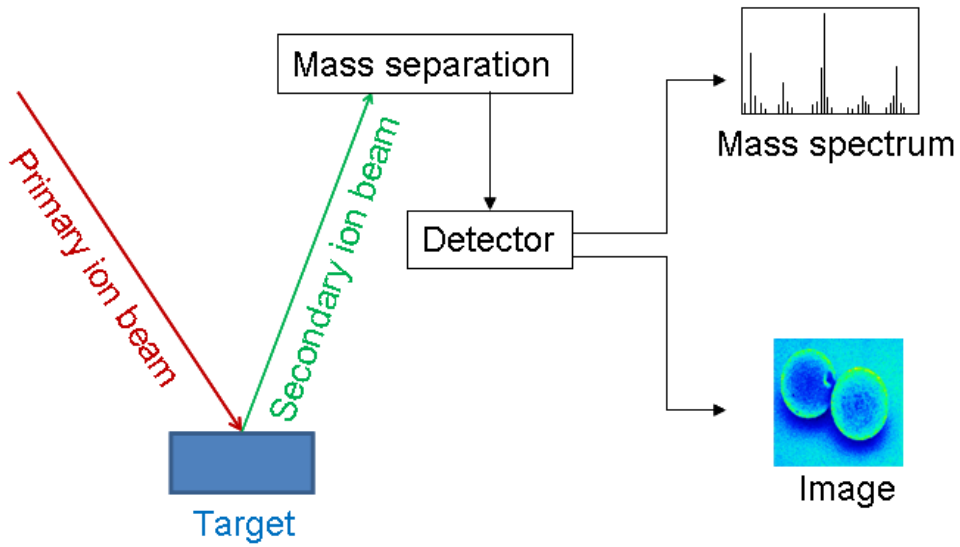

Figure 1.5: ToF-SIMS system. A beam of energetic (between 0.5 keV and 20 keV) primary ions (e.g., Au_3^+ or C_{60}^+) bombards the target surface and sputters atoms, molecules, and molecular fragments (secondary ions) from the target surface. The secondary ions ejected from the target surface are electrostatically accelerated into a "flight tube" and separated according to their m/z.

1.4 Classification

Classification is fundamental in scientific fields exploring empirical data. Data with similar characteristics can be grouped together to be better investigated. After developing unifying models explaining the occurrence of data, unknown data can be characteristically predicted.

There are three main categories of general classification methods: supervised classification (or simply classification), unsupervised classification (or clustering), and semi-supervised classification.

1. In supervised classification, a model or classifier is constructed to classify or predict unknown data into a known set of categories (classes) given data from the known set

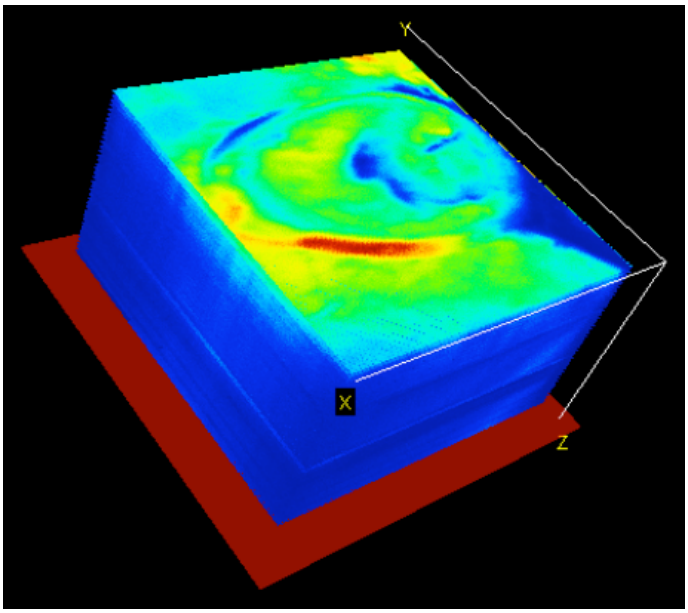

Figure 1.6: Three-dimensional ToF-SIMS TIC image of a frog oocyte. The x-axis, y-axis, and z-axis represent the location on target. The value of each pixel is the summation of all the intensity values of the mass spectrum associated with the pixel.

of categories.

2. In unsupervised classification, unknown data are grouped together such that data in the same group are in some sense more similar or homogeneous with one another than with data belonging to other groups [11].

3. In semi-supervised classification, a model or classifier is constructed to classify unknown data based on small amount of known data which may be difficult, expensive, or time consuming to obtain (e.g., known data may require the efforts of experienced human annotators), together with large amount of unknown data which are relatively easy to collect (e.g., unknown data may require less human effort).

This dissertation focuses on supervised classification. Classification in this dissertation refers to supervised classification.

1.4.1 Classification of Mass Spectra

Mass spectra contain characteristic information regarding the composition of compounds and properties of compounds. The mass spectrum of a compound can be used as a chemical "fingerprint" to characterize the compound [4]. The mass spectra of compounds from the same chemically related group are similar [12], *i.e.*, they may have similar sets of m/z values that have significant intensity values, but the intensity values may vary. Therefore, mass spectra can be used to predict or explain compound properties, such as biological or chemical properties, based on mass spectral similarity. Mass spectra have been used in diverse areas, including food and flavor analyses [13–15], studies of drug metabolism [16, 17], biomedical research [18, 19], environmental science [20], etc. As mass spectral data are used in more and more research areas, mass spectral analysis becomes more important.

Classification is one of the fundamental methodologies for analyzing mass spectral data. The primary goals of classification are to automatically group compounds based on their mass spectra, to find correlation between the properties of compounds and their mass spectra, and to provide a positive identification of unknown compounds.

Classification complements library search [21, 22] which searches a mass spectral library to identify unknown mass spectra. There are various mass spectral libraries, for example, the NIST/EPA/NIH Mass Spectral Library 2008 (NIST08) containing 220460 mass spectra. If an unknown compound is fairly common, its mass spectrum may be in a library, and correct identification of the compound through library search often is possible. If the unknown compound's mass spectrum is not present in a library, not only is the search result not a correct identification of the unknown compound, the search result often is misleading. For mass spectra that cannot be found in a library, classification can involve

identification of substructure types or well defined compound classes in order to establish and confirm structural conjectures or reveal relationships between mass spectra and chemical structures [23]. Classification also can be useful in cases when only structurally related compounds need to be retrieved.

Mass spectra are high-dimensional data. Mass spectra of complex mixtures are enormously complex with large mass ranges and many structurally significant peaks combined with noise peaks (such as contaminants and small or non-diagnostic fragment ions). Within this high-dimensional complexity, there is a huge amount of information about the identity of the mixture, e.g., compound composition, molecular orientation, surface order, and chemical bonding [24]. Establishing new mathematical or statistical methodologies for comprehensive information analysis and classification has become one of the most important tasks in mass spectral analysis.

1.4.2 Cross-Sample Classification of Comprehensive Two-Dimensional Chromatograms

Comprehensive separation and analysis of complex biological samples is a substantial challenge because of the presence of thousands of constituent compounds with highly variable concentrations and ranges and diverse physicochemical properties and detectability [25]. Two-dimensional separation patterns obtained by comprehensive chromatography, in particular GC×GC, analyze a complex mixture to characterize its composition. GC×GC is a powerful tool for complex biological sample characterization, differentiation, discrimination, and classification on the basis of the components distribution over the two-dimensional plane.

Comprehensive two-dimensional chromatography yields highly informative separation patterns because of its great practical peak capacity and sensitivity produced by applying two different separation principles (one for each chromatographic dimension). However, the improvement in information yields complex data (consisting of two-dimensional retention data and mass spectra) requiring comprehensive analyses to interpret the rich information and to extract useful information on sample characterization [26]. Cross-sample analysis of complex biological samples, such as sample classification, is even more challenging because of the difficulty of analyzing and interpreting the massive, complex data from many samples for relevant biochemical features. The large dimensionality of biological data, as well as the size of the dataset, and the possibility that significant chemical characteristics across many samples may be subtle and involve patterns of variations in multiple constituents, necessitate the investigation and development of new analysis methodologies.

1.5 Summary of Contributions

This dissertation presents a new supervised classification algorithm for classification of mass spectra, the most similar neighbor with a probability-based spectrum similarity measure (MSN-PSSM) [27] as described in Chapter 3. The MSN-PSSM algorithm is a multi-class classification algorithm that can deal with multiple classes directly without converting a multi-class problem into a set of two-class problems. The MSN-PSSM algorithm models the intra-class variability and uses a smoothing model in the similarity measure to enhance the robustness with respect to noise, such as chemical noise and instrument noise. The MSN-PSSM algorithm characterizes the domain information of labeled data by an array of probability distribution functions of intensities as a function of m/z. Each probability in the distribution function is the fraction of spectra in the labeled data having that inten-

sity value at the given m/z. The MSN-PSSM algorithm considers all m/z that contain discriminating information to avoid information loss.

Experimental results demonstrate the effectiveness and robustness of the new MSN-PSSM algorithm [27]. In leave-one-out cross-validation, it outperforms popular classification techniques for classification of mass spectra, such as principal component analysis (PCA) with discriminant function analysis (DFA), soft independent modeling of class analogy (SIMCA), and decision tree learning.

This dissertation also contributes to a new non-targeted cross-sample classification method to analyze comprehensive two-dimensional chromatograms [28, 29] as described in Chapter 5. The non-targeted cross-sample classification systematically and automatically detects registration peaks of multiple comprehensive two-dimensional chromatograms. Then, the non-targeted cross-sample classification uses the registration peaks to align (register) the chromatograms to generate a cumulative chromatogram. The registration peaks and the retention-time regions of all peaks detected in the cumulative chromatogram are used to generate a feature template. The registration peaks in the feature template are matched to the detected peaks in each chromatogram. For each chromatogram, the non-targeted cross-sample classification creates a feature vector that characterizes the detector response within the regions of the feature template. Then, the non-targeted cross-sample classification uses the feature vectors for the set of comprehensive two-dimensional chromatograms to perform classification and potential biomarker identification.

The new non-targeted cross-sample classification is successfully applied to a set of comprehensive two-dimensional chromatograms of breast cancer tumor samples, each from different individuals, for cancer grades 1 to 3 (as labeled by a cancer pathologist) [28]. Experimental results demonstrate the effectiveness of the new non-targeted cross-sample

classification. The feature vectors generated by the new non-targeted cross-sample classification are useful for discriminating between breast cancer tumor samples of different grades and providing information that can be used to identify potential biomarkers for closer examination.

1.6 Organization

This chapter briefly introduced the basic procedure of mass spectrometry, types of mass spectrometers, and nature of mass spectra. Then, this chapter introduced comprehensive two-dimensional gas chromatography with mass spectrometry and time-of-flight secondary ion mass spectrometry. Finally, this chapter described classification of mass spectra and cross-sample classification of comprehensive two-dimensional chromatograms.

Chapter 2 describes several popular supervised classification algorithms for mass spectra classification, including principal component analysis with discriminant function analysis, soft independent modeling of class analogy, and decision tree learning. Chapter 2 also describes research efforts to develop multivariate analysis techniques aimed at determining salient features of comprehensive two-dimensional chromatograms and quantitatively classifying complex samples.

Chapter 3 presents a new supervised classification algorithm, the most similar neighbor with a probability-based spectrum similarity measure. There are seven steps of this new algorithm: normalization, domain information characterization, intra-class variability model construction, smoothing model construction, probability-based spectrum similarity calculation, the most similar mass spectrum selection, and class label prediction.

Chapter 4 presents experimental results of the new supervised classification algorithm compared with popular classification techniques for mass spectra classification. The experimental results demonstrate the effectiveness and robustness of the new MSN-PSSM algorithm.

Chapter 5 presents a new non-targeted cross-sample classification method for comprehensive two-dimensional chromatograms. There are six steps of this new non-targeted cross-sample classification: chromatogram processing, registration template construction, cumulative chromatogram generation, feature template construction, cross-sample feature vector generation, and classification.

Chapter 6 presents experimental results of the new non-targeted cross-sample classification on a set of comprehensive two-dimensional chromatograms of breast cancer tumor samples. The experimental results demonstrate the effectiveness of the new non-targeted cross-sample classification.

Chapter 7 contains concluding remarks and ideas for future work.

Chapter 2

Related Work

This chapter describes several popular supervised classification algorithms for mass spectra classification. This chapter also describes research efforts to develop multivariate analysis techniques aimed at determining salient features of comprehensive two-dimensional chromatograms and quantitatively classifying complex samples.

2.1 Classification of Mass Spectra

Principal component analysis [30, 31] with discriminant function analysis [32, 33], soft independent modeling of class analogy [34], and decision tree learning [35] are popular supervised classification algorithms for classification of mass spectral data from GC$\times$GC-MS and ToF-SIMS.

2.1.1 Principal Component Analysis with Discriminant Function Analysis

Principal component analysis (PCA) with discriminant function analysis (DFA) is a supervised classification algorithm that is widely used for multivariate data analysis in comparing, discriminating, and classifying mass spectral data [36–41].

Principal component analysis was designed by Karl Pearson in 1901 [30]. The algorithm was introduced to psychologists in 1933 by Harold Hotelling, hence sometimes it is called Hotelling's transform [31]. Nowadays, it is mostly used as a tool in exploratory data analysis and for predictive modeling [42, 43].

PCA compares multiple mass spectra on the basis of multiple peaks in each mass spectrum. The multiple peaks are a large number of interrelated variables. The central idea of PCA in mass spectral analysis is to reduce the high dimensionality and simplify the mass spectra by transforming the large set of interrelated variables to a small set of uncorrelated variables called principal components that describe orthogonal directions of variance in the multiple mass spectra. The relationships of the multiple mass spectra are more apparent in the new coordinate system than in the original coordinates.

Given a mass spectral data matrix X with n rows (each row corresponds to a base-peak normalized mass spectrum) and l columns (each column corresponds to a m/z), PCA uses singular value decomposition (SVD) to decompose X into a score matrix Y times a loading matrix P^T plus a residual error matrix E, as Figure 2.1 illustrates:

$$X = y_1 p_1^T + y_2 p_2^T + \ldots + y_g p_g^T + E = Y P^T + E, \tag{2.1}$$

where:

- X is the mean centered data matrix;
- g is the dimension of the new space and $g \leq l$;
- E is the residual error matrix;
- the column vectors of P $\{p_i \mid i = 1, 2, \ldots, g\}$ are loadings (also called principal components or eigenvectors) which contain information on how variables relate to each other, and the $\{p_i\}$ vectors are orthonormal:

$$p_i^T p_j = \begin{cases} 0 & i \neq j \\ 1 & i = j; \end{cases} \tag{2.2}$$

- the column vectors of Y $\{y_i \mid i = 1, 2, \ldots, g\}$ are scores that are the projection of the mass spectra onto each principal component and contain information on how the mass spectra in the dataset relate to each other, and the score vector y_i is the linear combination of X defined by p_i:

$$Xp_i = y_i. \tag{2.3}$$

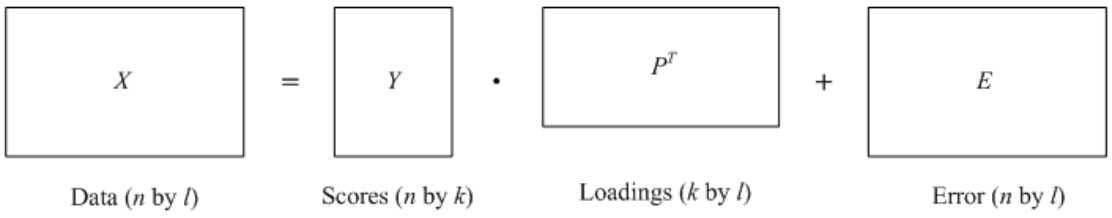

Figure 2.1: PCA decomposes a data matrix X into a score matrix Y times a loading matrix P^T plus a residual error matrix E.

The principal components are ranked in descending order by their eigenvalues. The eigenvalue is proportional to the amount of variance (information) captured by each eigenvector. The first principal component lies in the direction of most variance with subsequent principal components capturing less variance. The first few principal components retain most of the variation present in all of the original variables (peak intensities) and usually are adequate to approximate the original data. Therefore, the values in the remaining components may be dropped with minimal loss of information. In this manner, PCA is used for dimensionality reduction. One method to choose the number of principal components (g) to approximate the original data is selecting a cumulative percentage of total variation which one desires that the selected principal components contribute, for example, 80% or 90%. The cumulative percentage of total variation accounted for by the first g principal components is defined as:

$$q_g = \frac{100 \sum_{i=1}^{g} s_i}{\sum_{i=1}^{l} s_i}, \tag{2.4}$$

where:

- q_g represents the cumulative percentage of total variation accounted for by the first g ordered principal components;
- s_i represents the variance of the ith ordered principal component.

The required number of principal components is the smallest value of g for which the chosen percentage is exceeded. Jolliffe [42] discussed more methods for choosing the number of principal components.

PCA assumes that the data are linear combinations of the principal components ($X = YP^T$) and the original variables have a multivariate normal distribution (mean and variance

are sufficient statistics) [43, 44]. PCA also assumes that large variances reflect important information of the data and low variances reflect noise of the data. However, there is no guarantee that the directions of maximum variance provide good features for discrimination. It is widely recognized that the effectiveness of PCA also is dependent on appropriate data pretreatment, such as scaling, centering, removing contaminant peaks and background peaks from spectra, and non-linear transformations (for example the logarithm), but no definitive rules have been established for data pretreatment of PCA [37].

PCA, which is an unsupervised technique, reduces the dimensionality (from l to g) and simplifies the mass spectral data (from X to Y) whilst preserving most of the variance of the original data. Discriminant function analysis (DFA), which is a supervised technique, establishes classification models in the principal component space based on labeled mass spectra and uses classification models to classify unknown mass spectra. The inversion calculation of a covariance matrix in DFA makes it mathematically not applicable to highly interrelated variables [45]. Therefore, DFA is applied in the principal component space.

Given the score matrix Y (n rows and g columns) which is the projection of the mass spectral data matrix X in the principal component space, g principal component variables of the principal component space $p_1, p_2, \ldots, p_g$, and p predefined classes $\{c_j \mid j = 1, 2, \ldots, p\}$, DFA constructs linear combinations of $p_1, p_2, \ldots, p_g$ in such a way that the differences between the means of the predefined classes are maximized relative to the variance within classes:

$$Z_i = d_{i1}p_1 + d_{i2}p_2 + \ldots + d_{ig}p_g, \tag{2.5}$$

where:

- Z_i represents the ith linear combination of $p_1, p_2, \ldots, p_g$ and is referred to as the ith discriminant function;

- d_{i1} through d_{ig} represent discriminant coefficients of the ith discriminant function and maximize the difference between the means of the predefined classes;
- p_1 through p_g represent g principal component variables of the principal component space and are referred to as discriminant variables in DFA;
- g is the number of discriminant variables.

In a discriminant function, the discriminant variables with the largest discriminant coefficients are the ones that contribute most to the prediction of class membership. The value of a discriminant function resulting from applying the discriminant function to a given case of data is called the discriminant score of the case on the discriminant function. Figure 2.2 illustrates an example of applying two discriminant functions to one case of data with two discriminant variables. In Figure 2.2, Z_1 and Z_2 are discriminant functions, A is one case of data, p_1 and p_2 are discriminant variables, and c_1 through c_7 are seven predefined classes with class means represented by stars. The two discriminant functions are:

$$\begin{aligned} Z_1 &= d_{11}p_1 + d_{12}p_2, \\ Z_2 &= d_{21}p_1 + d_{22}p_2. \end{aligned} \tag{2.6}$$

For case A, the discriminant scores of the two discriminant functions are calculated as:

$$\begin{aligned} Z_{A1} &= d_{11}Y_{A1} + d_{12}Y_{A2}, \\ Z_{A2} &= d_{21}Y_{A1} + d_{22}Y_{A2}, \end{aligned} \tag{2.7}$$

where:

- Y_{A1} and Y_{A2} are the scores of case A in the p_1 and p_2 space (the projection values of case A onto the p_1 and p_2 space);

- Z_{A1} and Z_{A2} are the projection values of case A onto the Z_1 and Z_2 space and are referred to as the discriminant scores.

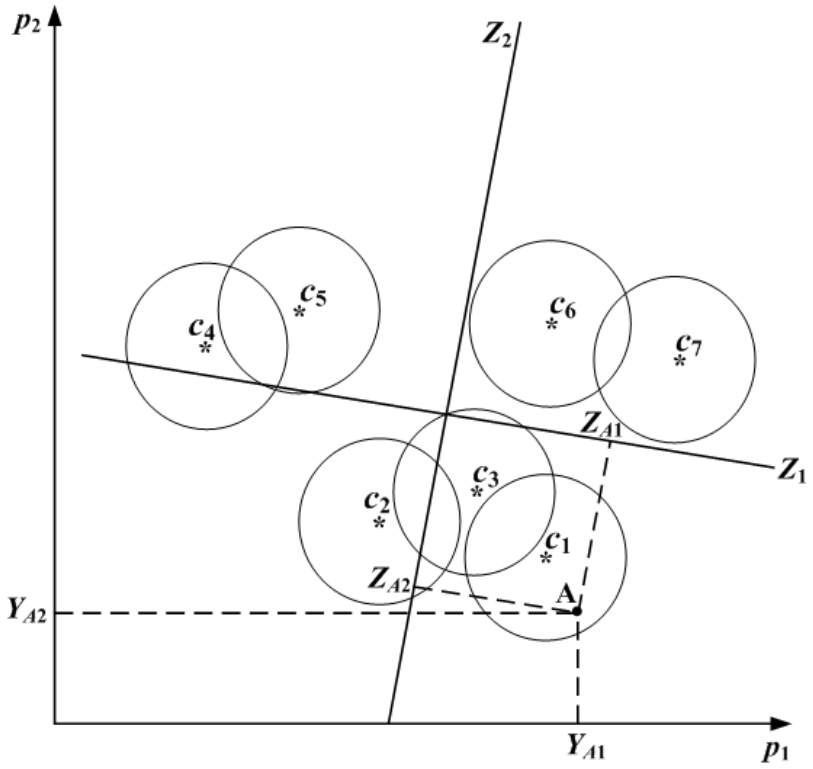

Figure 2.2: An example of applying two discriminant functions to one case of data with two discriminant variables. Z_1 and Z_2 are discriminant functions, A is one case of data, p_1 and p_2 are discriminant variables, and c_1 through c_7 are seven predefined classes with class means represented by stars.

DFA determines the discriminant coefficients of the discriminant functions by choosing the discriminant coefficients to maximize the F-ratios [32]:

$$F_i = \frac{d_i^T B d_i}{d_i^T W d_i}, \tag{2.8}$$

where:

- F_i represents the F-ratio of the ith discriminant function;

- $d_i = [d_{i1} \quad d_{i2} \quad \dots \quad d_{ig}]^T$ represents the discriminant coefficient vector of the ith discriminant function;

- W represents the within-group covariance matrix and is calculated as:

$$W = \frac{1}{n-p} \sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{ji} - \bar{y}_j)(y_{ji} - \bar{y}_j)^T; \tag{2.9}$$

- B represents the between-group covariance matrix and is calculated as:

$$B = \frac{1}{p-1} \sum_{j=1}^{p} n_j (\bar{y}_j - \bar{y})(\bar{y}_j - \bar{y})^T. \tag{2.10}$$

In Equation (2.9) and Equation (2.10),

- n represents the number of data;

- p represents the number of predefined classes;

- $n_j (j = 1, 2, \dots, p)$ represents the number of data in each class, and $n = \sum_{j=1}^{p} n_j$;

- y_{ji} is a column vector and represents the projection of the ith data case of the jth class onto the principal component space;

- $\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ji}$ is a column vector and represents the mean of the y_{ji} in the jth class;

- $\bar{y} = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n_j} y_{ji}$ is a column vector and represents the mean of the y_{ji} in the whole data set.

The within-group covariance matrix reflects the squared deviations and cross deviations of each data case from the mean of its class. The between-group covariance matrix reflects the squared deviations and cross deviations of each class mean from the mean of

the class means. The larger the F-ratio, the more differences between classes than within classes. The number of discriminant functions h is equal to the number of classes minus one, or the number of discriminant variables in the analysis, whichever is smaller. The first discriminant function Z_1 has the largest F-ratio and provides the most overall discrimination between classes. The second discriminant function Z_2 has the second largest F-ratio and captures as much as possible of the class differences not captured by Z_1. And so on. Finding the discriminant coefficients that maximize the F-ratio is an eigenvalue problem. The eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_h > 0$ of the matrix $W^{-1}B$ provide the F-ratios of $Z_1, Z_2, \ldots, Z_h$, and the corresponding eigenvectors provide the discriminant coefficient vectors of $Z_1, Z_2, \ldots, Z_h$.

To classify unknown data, DFA establishes a classification function for each class given p predefined classes [46]:

$$C_j = c_{j0} + c_{j1}p_1 + c_{j2}p_2 + \ldots + c_{jg}p_g, \tag{2.11}$$

where:

- C_j represents the classification function of class j, and $j = 1, 2, \ldots, p$;
- p_1 through p_g represent g discriminant variables;
- g is the number of discriminant variables;
- c_{j1} through c_{jg} represent the classification coefficients of class j and

$$[c_{j1} \quad c_{j2} \quad \ldots \quad c_{jg}]^T = W^{-1}\bar{y}_j, \tag{2.12}$$

  where W is the within-group covariance matrix and $\bar{y}_j$ (a column vector) is the mean of class j;

- c_{j0} represents a constant, and

$$c_{j0} = -\frac{1}{2}[c_{j1} \quad c_{j2} \quad \dots \quad c_{jg}]\bar{y}_j. \tag{2.13}$$

The value of a classification function resulting from applying the classification function to an unknown case of data is called the classification score of the case on the classification function. Each unknown case of data has p classification scores, one for each class. An unknown case of data is classified to the class for which it has the highest classification score.

DFA assumes that the discriminant variables are not completely redundant. If any one of the variables is completely redundant with the other variables then the within-group covariance matrix is ill-conditioned (singular) and cannot be inverted to construct discriminant functions [47, 48]. DFA is applied in the principal component space in this study, which guarantees that this assumption is satisfied. DFA also assumes that the discriminant variables are multivariate normally distributed within classes [33]. For the projections of mass spectra data in the principal space, this assumption is satisfied. Another assumption of DFA is that the within-group covariance matrix is the same for all classes [33] which is not necessarily true for the projections of mass spectra data in the principal space. Lindman [49] showed that the F-ratio is quite robust against violations of this assumption, and minor violations of this assumption is usually not fatal [50].

2.1.2 Soft Independent Modeling of Class Analogy

Soft independent modeling of class analogy (SIMCA) is a supervised classification algorithm proposed by Svante Wold and Michael Sjostrom in 1976 [34]. Some modified algo-

rithms were proposed thereafter [51–54]. SIMCA is another popular multivariate data analysis technique in comparing, discriminating, and classifying mass spectral data [37, 55–59].

SIMCA describes the multivariate data structure of each predefined class of mass spectra separately in a reduced space using PCA. The unknown mass spectra are classified to the established class model with the minimum orthogonal distance.

Given p predefined classes of mass spectra, denote each mass spectral class by $X_j (j = 1, 2, \ldots, p)$ with n_j rows (each row corresponds to a base-peak normalized mass spectrum) and l columns (each column corresponds to a m/z). A sufficient number of principal components are retained to account for most of the variation within each class following Equation (2.1):

$$X_j = Y_j P_j^T + E_j, \tag{2.14}$$

where:

- P_j represents the loading (also called the principal component or eigenvector) matrix of class j, and each column of P_j represents one principal component;
- Y_j represents the score matrix of class j which is the projection of mass spectra onto each principal component;
- E_j represents the residual error matrix of class j.

The variance explained by the principal component models is called the modeled variance and the variance not accounted for by the principal component models is called the residual variance.

After each predefined class is represented by a principal component model, an unknown

base-peak normalized mass spectrum $x = [x_1 \quad x_2 \quad \dots \quad x_l]^T$ is projected onto each class's model to yield a score vector y_j of x in each class's PCA space:

$$y_j = P_j^T x. \tag{2.15}$$

Then, the score y_j is back-projected to the original space to yield an estimation $\widehat{x}_j = [\widehat{x}_{j1} \quad \widehat{x}_{j2} \quad \dots \quad \widehat{x}_{jl}]^T$ of x with residual variance dropped:

$$\widehat{x}_j = P_j y_j. \tag{2.16}$$

SIMCA defines the orthogonal distance of the unknown base-peak normalized mass spectrum x to the PCA space of class j as [45]:

$$D_j = \sqrt{\sum_{i=1}^{l} (x_i - \widehat{x}_{ji})^2}. \tag{2.17}$$

The orthogonal distance measures the deviation of a mass spectrum to a class's PCA model and is used as a similarity measure. The smaller the orthogonal distance between a mass spectrum and the PCA space of a class, the more similar the mass spectrum is to that class. The unknown mass spectrum is classified to the established class model with the minimum orthogonal distance.

SIMCA models each class with a separate PCA model, and the assumptions of SIMCA are the same as the assumptions of PCA. A known problem of SIMCA is the sensitivity to the quality of the data used to establish the principal component models. SIMCA is successful if the predefined classes form compact groups.

2.1.3 Decision Tree Learning

Decision tree learning is a supervised classification algorithm commonly used to compare, discriminate, and classify mass spectral data [60–66]. Decision tree learning constructs a classification model represented by a tree based on known data. Each node in the decision tree specifies a variable test of the known data. A selection measure (e.g., the information gain measure [67], the gain ratio measure [67], and the Gini index measure [68]) selects the variable test. Each branch descending from each node corresponds to one of the possible test results for the variable. A decision tree is learned by splitting the known data into subsets based on a variable test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the known data of a subset at a node are all in the same class. Unknown data are sorted down the decision tree from the root to one of the leaf nodes, which provides the predicted class. In each node, a variable test selection measure evaluates how well each variable test splits the subset into further subsets such that known data generally are in the same class and selects the best variable test to further split the subset. Information gain, gain ratio, and Gini index are three commonly used variable test selection measures.

Information gain is based on the concept of entropy from information theory. Given p predefined classes $\{c_j \mid j = 1, 2, \ldots, p\}$, a set X of known data, and a variable test t that splits X to subsets $X_i (i = 1, 2, \ldots, a)$, information gain is [67]:

$$gain(X, t) = info(X) - info_t(X), \tag{2.18}$$

where:

- t represents a variable test;

- X is a set of known data;
- $gain(X, t)$ is the information that is gained by branching on variable test t;
- $info(X)$ is the information before branching on variable test t;
- $info_t(X)$ is the information after branching on variable test t.

In Equation (2.18), $info(X)$ measures the average amount of information needed to identify the class of a datum in X. The function $info(X)$ is also known as the entropy of X and is calculated as:

$$info(X) = -\sum_{j=1}^{p} \frac{freq(c_j, X)}{|X|} \log_2 \frac{freq(c_j, X)}{|X|}, \tag{2.19}$$

where:

- $freq(c_j, X)$ is the number of data items in set X that belong to class c_j;
- $|X|$ is the number of data items in set X.

In Equation (2.18), $info_t(X)$ measures the information after X has been split to the a subsets $X_i (i = 1, 2, \ldots, a)$ in accordance with variable test t:

$$info_t(X) = \sum_{i=1}^{a} \frac{|X_i|}{|X|} info(X_i), \tag{2.20}$$

where $|X_i|$ represents the number of data items in set X_i. Information gain measures the information that is gained by splitting X in accordance with the variable test t. The information gain measure selects the variable test with maximum information gain in each node.

Information gain has a natural bias on variables with many values and may generate broad decision trees of small depth. Gain ratio suppresses this bias by a normalization.

Gain ratio is defined as [67]:

$$gain\_ratio(X,t) = \frac{gain(X,t)}{split\_info(X,t)}, \tag{2.21}$$

where $split\_info(X,t)$ represents the potential information generated by splitting X into a subsets $X_i (i = 1, 2, \ldots, a)$, and is defined as:

$$split\_info(t) = -\sum_{i=1}^{a} \frac{|X_i|}{|X|} \log_2 \frac{|X_i|}{|X|}. \tag{2.22}$$

In Equation (2.21), $gain\_ratio(X,t)$ expresses the proportion of information generated by the split that appears helpful for classification. If the split is near trivial, split information will be small and this ratio will be unstable. To avoid this, the gain ratio measure selects a test to maximize the gain ratio, subject to the constraint that the information gain must be large (at least as great as the average gain over all tests examined). The gain ratio measure tends to prefer unbalanced splits in which one subset is much smaller than the others.

The Gini index considers a binary split for each node. Given p predefined classes $\{c_j \mid j = 1, 2, \ldots, p\}$, a set X of known data, and a binary split on t which splits set X into subsets X_1 and X_2, the reduction in impurity that would be incurred by the split on variable test t is [68]:

$$reduction(X,t) = Gini(X) - Gini_t(X), \tag{2.23}$$

where:

- t represents a variable test;
- X is a set of known data;

- $reduction(X, t)$ is the reduction in impurity that would be incurred by the binary split on variable test t;
- $Gini(X)$ is the Gini index before splitting on variable test t;
- $Gini_t(X)$ is the Gini index after splitting on variable test t.

In Equation (2.23), $Gini(X)$ is based on squared probabilities of membership for each class in the node, and is calculated as:

$$Gini(X) = 1 - \sum_{j=1}^{p} \left(\frac{|X_j|}{|X|} \right)^2, \tag{2.24}$$

where:

- $Gini(X)$ is the Gini index, which measures the impurity of X;
- $|X_j|$ is the number of data items in set X belonging to class c_j;
- $|X|$ is the number of data items in set X.

In Equation (2.23), $Gini_t(X)$ measures the Gini index after splitting on variable test t, and is computed as a weighted sum of the impurity of each resulting subset:

$$Gini_t(X) = \frac{|X_1|}{|X|} Gini(X_1) + \frac{|X_2|}{|X|} Gini(X_2). \tag{2.25}$$

The reduction in impurity that would be incurred by a binary split reaches its minimum (zero) when all data in the node fall into a single class. The Gini index selects the variable test with the largest reduction in impurity. The Gini index tends to favor equal sized splittings.

Decision tree learning has an intuitive representation and the resulting model is easy to understand. Moreover, the decision tree provides a nonparametric model, and no intervention is required from the user. Thus, decision tree learning is suited for exploratory knowledge discovery. One major problem with decision trees is their high volatility [69]. A small change in the data set often can result in a very different series of splits, which make later interpretation somewhat precarious and difficult.

2.2 Cross-Sample Classification of Comprehensive Two-Dimensional Chromatograms

Targeted analysis and non-targeted analysis are two general approaches to extract consistent information and analyze comprehensive two-dimensional chromatograms.

In targeted analysis, only a few specific compounds are identified and compared from chromatogram to chromatogram or only certain regions of the comprehensive two-dimensional chromatograms are desired for their representative characterization of chromatograms. In targeted analysis, analysts are involved in selecting of target compounds or regions. Targeted analysis uses limited information from comprehensive two-dimensional chromatograms, which could miss important information in the non-selected compounds or regions.

In non-targeted analysis, all compounds of comprehensive two-dimensional chromatograms are important to provide comprehensive surveys of qualitative and quantitative differences in the chemical composition between chromatograms, prompting research efforts to develop information extraction methods and multivariate analysis techniques to determine salient features and construct discriminant models of comprehensive two-dimensional chro-

matograms.

Researchers have applied varied approaches on non-targeted analysis: visual image comparisons, data point comparisons, region comparisons, peak comparisons, and peak-based region comparisons to extract consistent information and discriminate comprehensive two-dimensional chromatograms.

2.2.1 Visual Image Comparisons

Visual image comparisons are primarily qualitative visual comparisons of chromatograms without benefit of software designed for operating on comprehensive two-dimensional chromatograms.

Blomberg *et al*. [70] analyzed and compared GC×GC chromatograms of a distillation cut of a heavy oil and its hydrogenated product to illustrate the conversion of olefins and sulphur compounds. Similarly, Gaines *et al*. [71], Reddy *et al*. [72], and Reddy *et al*. [73] analyzed and compared GC×GC chromatograms of different oil samples.

Perera *et al*. [74] analyzed GC×GC chromatograms of emissions of volatile organic compounds from mechanically wounded plants to identify and compare the major chemical species emitted from plants after mechanical wounding.

Janssen *et al*. [75] analyzed and compared comprehensive two-dimensional liquid chromatography × gas chromatography (LC×GC) chromatograms of edible oils and fats in triglycerides classification.

Qualitative visualizations do not provide quantitative comparisons and are insufficient

for recognizing subtle differences.

2.2.2 Data Point Comparisons

In data point comparisons, chromatograms are compared point-by-point (or pixel-by-pixel).

Johnson and Synovec [76] utilized an analysis of variance (ANOVA)-based feature selection to identify chromatographic features and PCA to classify GC×GC chromatograms of jet fuel mixtures. The chromatographic features were generated by point-by-point ANOVA calculations, which provided a ratio for each retention time in the chromatograms. The retention times with a ratio greater than a selected threshold were extracted from chromatograms and analyzed by PCA for fuel type classification.

Shellie *et al*. [77] directly compared comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry (GC×GC-TOFMS) chromatograms of metabolite profiles of mouse tissue extracts using different methods including direct chromatogram subtraction, averaging routines, weighting factors, and t-test.

Pierce *et al*. [78] utilized a PCA based method to discover chemical differences in GC×GC-TOFMS chromatograms of metabolites in plant samples. The GC×GC-TOFMS produced a three-dimensional array of data for each sample (retention times on two chromatographic columns and a complete mass spectrum at every point in the separation space). This three-dimensional array was reduced to a two-dimensional matrix by using only one mass channel at a time to produce a *m/z* chromatogram. Then, PCA was applied to selected *m/z* chromatograms to compare the metabolite profiles of a set of different plant samples. Similarly, Mohler *et al*. [79] utilized PCA followed by parallel factor analysis (PARAFAC)

on selected *m/z* chromatograms to identify chemical differences in GC×GC-TOFMS chromatograms of metabolite extracts isolated from yeast cells grown under different conditions.

Pierce *et al*. [80] utilized a fisher ratio method to discover chemical differences in GC×GC-TOFMS separations of urine metabolite samples. A set of GC×GC-TOFMS sample profiles produced a four-dimensional array of data (the fourth dimension was the sample replicate dimension). This four-dimensional array was reduced to a two-dimensional array by a novel indexing scheme. The fisher ratio method calculated a fisher ratio at every point in this two-dimensional array to discover significant chemical differences between complex samples. Further, Mohler *et al*. [81] and Mohler *et al*. [82] combined a statistically-based fisher ratio method utilizing all mass channels with PARAFAC to compare and analyze GC×GC-TOFMS chromatograms of yeast metabolites. Similarly, Guo and Lidstrom [83] applied fisher ratio analysis on selective *m/z* chromatograms with PARAFAC to identify the metabolite differences between cells grown on methanol and succinate.

Hollingsworth *et al*. [84] developed software methods for aligning chromatograms based on marker peaks and for comparative visualizing the point-by-point differences of GC×GC by various methods (for example, time-loop flicker and colourization). Subsequently, Almstetter *et al*. [85] developed retention time correction and data alignment tools to compare GC×GC-TOFMS metabolite profiles of a wild-type and a mutant strain of *Escherichia coli*.

Data point comparisons require precise chromatographic alignment, which is difficult over large sample sets.

2.2.3 Region Comparisons

Regions characterize multiple data points within each chromatogram (e.g., summing the responses at all data points in each region). In region comparisons, chromatograms are compared region-by-region.

Mispelaar [86] created a mesh of contiguous hand-drawn polygons to subjectively encompass different groups of interest in diesel samples and demonstrated the use of geometric transformations to better fit different GC×GC chromatograms.

Arey *et al*. [87] discretized GC×GC chromatograms into low-resolution, two-dimensional grids of cells to investigate weathered oil samples. Cell boundaries were defined by computed contours of hydrocarbon vapor pressure and aqueous solubility. Cell mass was defined as the sum of all pixel heights assigned to a cell. Mass loss tables were constructed to depict compositional changes along systematic coordinates of volatility and aqueous solubility. To mitigate the effect of misalignment, trapezoidal weighting functions were used at the borders between regions.

Rathbun *et al*. [88] used an enhanced template-based method to compare GC×GC chromatograms of complex petroleum samples and to identify chemical compounds. The enhanced template-based method constructed a series of contiguous retention-plane regions (area objects) as features of a chromatogram. To compare the chromatograms, a region template recording all regions was constructed to match the petroleum chromatograms.

In region comparisons, selectivity is reduced to the extent that peaks of multiple analytes are included in the same region. This is especially problematic if a salient trace analyte is in the same region as a non-salient predominant analyte.

2.2.4 Peak Comparisons

In comprehensive two-dimensional chromatograms, each individual chemical compound forms a two-dimensional cluster of pixels (a peak) with values larger than the background values (the data values in which no chemical peak is present). In peak comparisons, chromatograms are compared peak-by-peak.

Gaines *et al*. [71] provided an early demonstration of using quantitative characterizations of individual peaks and groups of peaks detected in GC×GC chromatograms to fingerprint samples of an oil spill and potential sources to identify the source of the spill.

Mispelaar [86, 89] used principal component discriminant analysis on peaks detected in GC×GC chromatograms to distinguish samples from different oil reservoirs. More than 6000 peaks were detected and filtered by time-based filtering and alignment checks, relative standard deviations, and a manual selection.

Porter *et al*. [90] used PARAFAC with alternating least squares and alternating least squares with flexible constraints on peaks detected in comprehensive two-dimensional liquid chromatography (LC×LC) chromatograms to discriminate mutant and wild-type maize samples.

Qui *et al*. [91] utilized PCA on individual peaks detected in GC×GC chromatograms to classify traditional Chinese medicine volatile oil samples from different geographical origins.

Oh *et al*. [92] developed a novel peak sorting algorithm (MSsort) to find metabolite peaks generated from the same metabolite but detected in different GC×GC-TOFMS chro-

matograms of human serum samples.

Gaquerel *et al*. [93] used ANOVA, hierarchical clustering analysis, and PCA on individual peaks detected in GC×GC-TOFMS chromatograms to compare volatile bouquets emitted after insect herbivory.

Li *et al*. [94] utilized orthogonal signal correction filtered partial least-squares discriminant analysis on peaks detected in GC×GC-TOFMS chromatograms to compare human plasma from diabetic patients and healthy controls and discover metabolites with a significant concentration change in diabetic patients.

Tan *et al*. [95] used PCA and partial least-square discriminant analysis to characterize and differentiate GC×GC-TOFMS chromatograms of two Chinese herbs.

Koek *et al*. [96] assessed the feasibility of using a processing strategy based on commercially available software for the unbiased, non-target automated quantification of as many metabolites as possible in mouse liver samples measured with GC×GC-MS.

In peak comparisons, peak matching is a critical challenge. Peak detection errors as well as the inherent ambiguity of matching both contribute to make comprehensive peak matching (i.e., matching all peaks) across many samples intractable.

2.2.5 Peak-Based Region Comparisons

In peak-based region comparisons, regions are defined for individual peaks and chromatograms are compared by peak-based regions.

Schmarr and Bernhardt [97] described an approach originating from the proteomics field to compare GC×GC chromatograms of volatile patterns from fruits. GC×GC chromatograms were analyzed utilizing a workflow derived from two-dimensional gel-based proteomics. Run-to-run variations among chromatograms were compensated by warping. The chromatograms were then merged into a fusion chromatogram yielding a project-wide peak consensus pattern. Within detected peak boundaries (regions) of this consensus pattern, relative quantities of the volatiles from each chromatogram were calculated. These profiles were used for multivariate statistical analysis and allowed clustering of comparable sample origins and prediction of unknown samples.

Peak-based region comparisons are more comprehensive than peak comparisons and are more selective than region comparisons. Alignment is a potential source of errors for peak-based region comparisons. Peak detection errors, such as unseparated coelutions and incorrectly split peaks, are another source of errors.

Chapter 3

The Most Similar Neighbor with a Probability-Based Spectrum Similarity Measure

This chapter presents a new supervised classification algorithm, the most similar neighbor with a probability-based spectrum similarity measure (MSN-PSSM) [27]. Given:

1. n labeled mass spectra $\{x_i \mid i = 1, 2, \ldots, n\}$, where x_i represents the vector of intensities of mass spectrum i;
2. p predefined classes $\{c_j \mid j = 1, 2, \ldots, p\}$, where c_j represents the class name of class j;
3. class labels of the n mass spectra $\{y_i \mid i = 1, 2, \ldots, n\}$, where $y_i \in \{c_j \mid j = 1, 2, \ldots, p\}$;

4. a query mass spectrum x_q, and $x_q \notin \{x_i \mid i = 1, 2, \ldots, n\}$;

the steps of the MSN-PSSM algorithm are:

1. Normalize the n labeled mass spectra $\{x_i \mid i = 1, 2, \ldots, n\}$ to $\{\bar{x}_i \mid i = 1, 2, \ldots, n\}$.
2. Express domain characteristics based on the normalized labeled data.
3. Build intra-class variability models for each predefined class c_j based on the normalized labeled data and their class labels.
4. Build smoothing models for each predefined class c_j based on the domain characteristics and the intra-class variability model for c_j.
5. Calculate the probability-based spectrum similarity between the normalize query spectrum $\bar{x}_q$ and each labeled datum $\bar{x}_i$.
6. Select the most similar datum of the query datum $\bar{x}_q$, *i.e.*, having the highest probability-based spectrum similarity with $\bar{x}_q$.
7. Predict the class label of the query datum $\bar{x}_q$ to be the most similar datum's class label.

The MSN-PSSM algorithm is a multi-class classification algorithm, that is, it can deal with multiple classes directly without converting a multi-class problem into a set of two-class problems. The MSN-PSSM algorithm models the intra-class variability and uses a smoothing model in the similarity measure to enhance the robustness with respect to noise, such as chemical noise and instrument noise. Some popular classification techniques (for example SIMCA) also consider the variance of each class in class modeling. Some other

popular classification techniques, such as PCA with DFA and decision tree learning, do not explicitly include the inter-class variability in their models. The MSN-PSSM algorithm characterizes the domain information of labeled data by an array of probability distribution functions of intensities as a function of m/z. Each probability in the distribution function is the fraction of spectra in the labeled data having that intensity value at the given m/z. The MSN-PSSM algorithm considers all m/z that contain discriminating information to avoid information loss. Some popular classification techniques, such as PCA with DFA, SIMCA, and decision tree learning, are highly selective in choosing the m/z and may lose useful discriminating information.

Figure 3.1 illustrates the MSN-PSSM algorithm step by step. The following sections describe the MSN-PSSM algorithm step-by-step in detail.

3.1 Normalization

When comparing mass spectra of different samples, the raw data (the ion intensities of various m/z peaks) cannot be used directly. The volumes and concentrations of samples are different due to various physical and chemical factors leading to different ion intensities. To compensate for the variations in samples, each mass spectrum is normalized. Base-peak normalization is a common normalization method in mass spectrometry to exclude sample volume fluctuation.

In base-peak normalization, the intensity of each peak in each labeled mass spectrum of $\{x_i \mid i = 1, 2, \ldots, n\}$ is normalized to the intensity of the most intense peak in that spectrum and rounded to the closest integer value, such that the intensity of the base peak

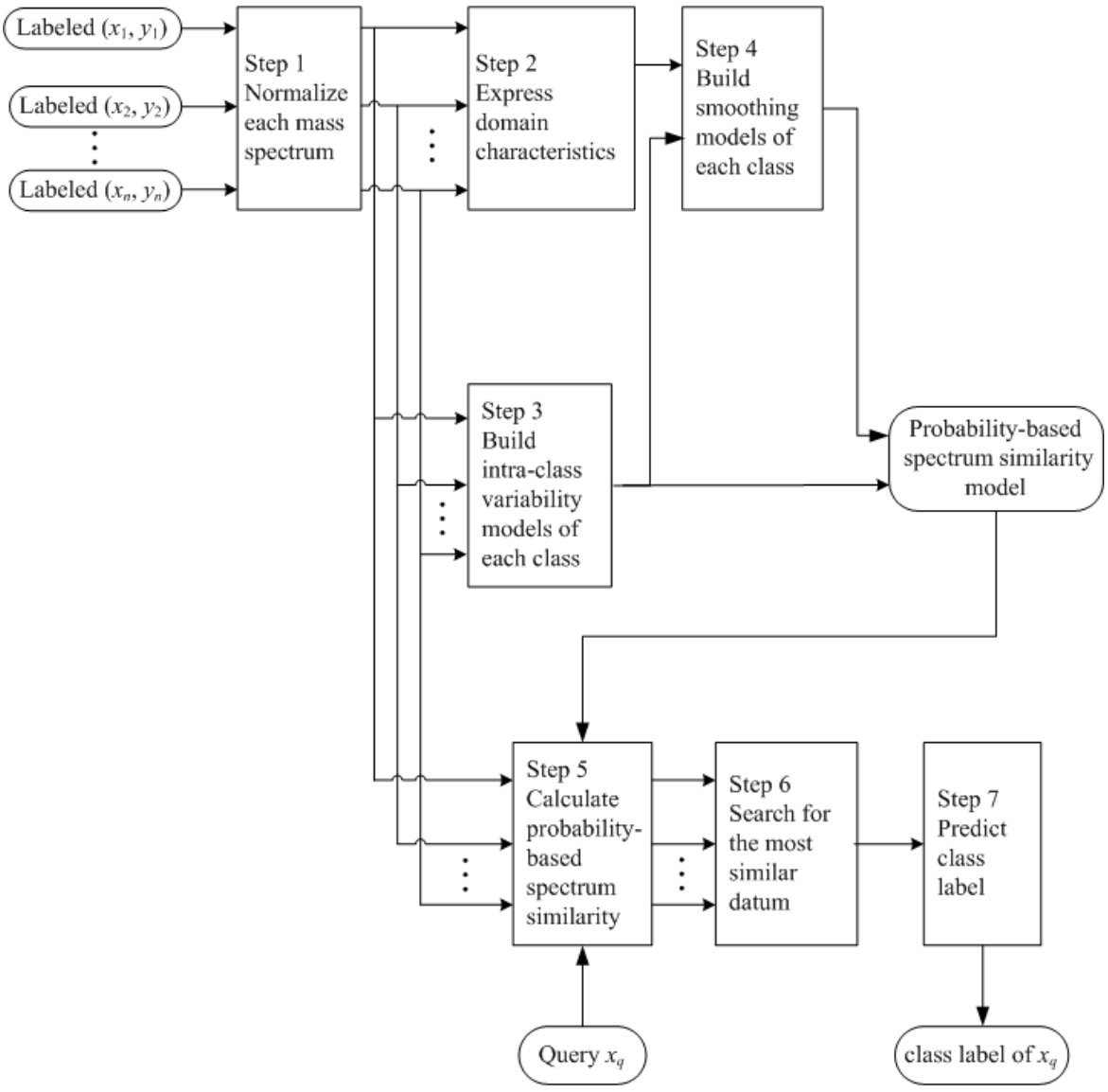

Figure 3.1: The steps of the MSN-PSSM algorithm.

is 999:

$$\bar{x}_i[m] = round\left(x_i[m] \times \frac{999}{x_i[m_b]}\right), \tag{3.1}$$

where:

- m is the mass-to-charge ratio;
- $\bar{x}_i[m]$ is the normalized intensity of mass spectrum x_i at mass-to-charge ratio m;
- $x_i[m]$ is the original intensity of mass spectrum x_i at mass-to-charge ratio m;
- $x_i[m_b]$ is the original intensity of mass spectrum x_i at the base peak base m_b, where $\forall m$, $x_i[m_b] \geq x_i[m]$.

3.2 Domain Characteristics

For the n normalized labeled data $\{\bar{x}_i \mid i = 1, 2, \ldots, n\}$, the intensity probability distribution at mass-to-charge ratio m is:

$$P_m[f] = \frac{n_m[f]}{n}, \tag{3.2}$$

where:

- f is the intensity, and $0 \leqslant f \leqslant 999$;
- n is the number of normalized mass spectra $\bar{x}_i$;
- $n_m[f]$ represents the number of normalized mass spectra $\bar{x}_i$ having intensity f at mass-to-charge ratio m;
- $P_m[f]$ represents the intensity probability of f at mass-to-charge ratio m.

The n normalized labeled data $\{\bar{x}_i \mid i = 1, 2, \ldots, n\}$ are characterized by the intensity probability distributions of all intensities f from 0 to 999 at all mass-to-charge ratio m from m_{min} to m_{max}. Each intensity probability in the distribution is the fraction of mass spectra in the domain having that intensity value at the given m/z. The intensity probability distributions provide the relation between intensity probabilities and m/z. The intensity probability distributions also provide the relation between intensity probabilities and intensities.

To simplify the domain characterization, m/z independence is assumed, that is, for a spectrum, the intensity value at one m/z is independent of the intensity value at any other m/z. This assumption is not usually true. Classification performance could be enhanced by considering inter-m/z dependence as discussed in Chapter 7. Establishing mathematical models for inter-m/z dependence is a substantial task and beyond the scope of this research.

3.3 Intra-Class Variability Model

The intra-class variability is modeled by statistically analyzing the data for a class. For the class c_j at mass-to-charge ratio m, the intensities are modeled as a Gaussian distribution:

$$\mathcal{N}_{j,m}[f] = \int_{f-\frac{1}{2}}^{f+\frac{1}{2}} \frac{1}{\sqrt{2\pi}\,\sigma_j[m]} \exp\left(\frac{-(x-\mu_j[m])^2}{2\sigma_j^2[m]}\right) dx, \tag{3.3}$$

where:

- $\mu_j[m]$ represents the mean of intensities at mass-to-charge ratio m for class c_j;
- $\sigma_j[m]$ represents the standard deviation of intensities at mass-to-charge ratio m for

class c_j;

- f is the intensity;
- $\mathcal{N}_{j,m}[f]$ represents the probability of intensity f at mass-to-charge ratio m for class c_j.

3.3.1 Parameter Estimation

The standard deviation of intensities at mass-to-charge ratio m tends to be intensity dependent, that is, the larger the intensity level at mass-to-charge ratio m, the greater the standard deviation at mass-to-charge ratio m. For the class c_j at mass-to-charge ratio m, the standard deviation of intensities is modeled as:

$$\sigma_j[m] = a_{j,m}\mu_j[m] + b_{j,m}, \tag{3.4}$$

where:

- $\sigma_j[m]$ is the standard deviation of intensities at mass-to-charge ratio m for class c_j;
- $\mu_j[m]$ is the mean of intensities at mass-to-charge ratio m for class c_j;
- $a_{j,m}$ and $b_{j,m}$ are linear regression parameters of standard deviations against intensity levels at mass-to-charge ratio m for class c_j.

The relationship between the standard deviation of intensities and the intensity level is assumed linear. A non-linear relationship may yield better results as discussed in Chapter 7.

Figure 3.2 illustrates the steps of parameter estimation of standard deviation of the intra-class variability model.

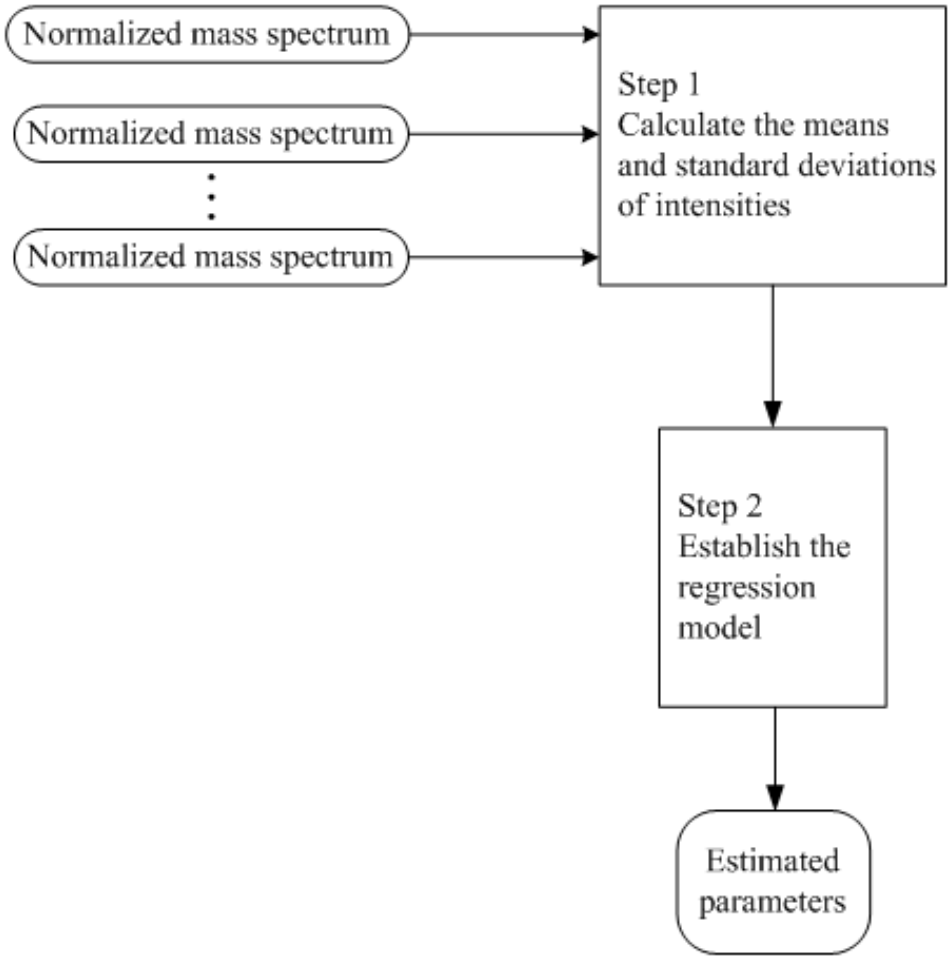

Figure 3.2: The steps of parameter estimation of standard deviation of the intra-class variability model.

The first step of parameter estimation is calculating the means and standard deviations of intensities. For base-peak normalized mass spectra of class c_j $\{(\bar{x}_i, y_i) \mid y_i = c_j\}$, the

average intensity of the normalized mass spectra at mass-to-charge ratio m is:

$$\mu_j[m] = \frac{1}{n_j} \sum_{y_i = c_j} \bar{x}_i[m], \tag{3.5}$$

where:

- $\mu_j[m]$ is the average intensity of the normalized mass spectra of class c_j at mass-to-charge ratio m;
- n_j is the number of mass spectra $\bar{x}_i$ with class label $y_i = c_j$;
- $\bar{x}_i[m]$ is the base-peak normalized intensity of mass spectrum x_i at mass-to-charge ratio m.

The standard deviation of intensities from the average intensity of the normalized mass spectra of class c_j at mass-to-charge ratio m is:

$$\sigma_j[m] = \sqrt{\frac{1}{n_j - 1} \sum_{y_i = c_j} [\bar{x}_i[m] - \mu_j[m]]^2}, \tag{3.6}$$

where:

- $\sigma_j[m]$ is the standard deviation of intensities of class c_j at mass-to-charge ratio m;
- n_j is the number of mass spectra $\bar{x}_i$ with class label $y_i = c_j$;
- $\bar{x}_i[m]$ is the base-peak normalized intensity of mass spectrum x_i at mass-to-charge ratio m;
- $\mu_j[m]$ is the average intensity of the normalized mass spectra of class c_j at mass-to-charge ratio m.

The second step of parameter estimation is establishing the regression model. The standard deviation of intensities of class c_j at mass-to-charge ratio m is modeled to be proportional to the intensity level at mass-to-charge ratio m of class c_j. Linear regression of the standard deviations over the average intensities of the normalized mass spectra yields the linear regression parameters as:

$$a_{j,m} = \frac{d \sum\limits_{\mu_j[m] \geq f_0} \mu_j[m]\sigma_j[m] - \sum\limits_{\mu_j[m] \geq f_0} \mu_j[m] \sum\limits_{\mu_j[m] \geq f_0} \sigma_j[m]}{d \sum\limits_{\mu_j[m] \geq f_0} \mu_j^2[m] - \left(\sum\limits_{\mu_j[m] \geq f_0} \mu_j[m] \right)^2}, \tag{3.7}$$

$$b_{j,m} = \frac{\sum\limits_{\mu_j[m] \geq f_0} \sigma_j[m] - a_{j,m} \sum\limits_{\mu_j[m] \geq f_0} \mu_j[m]}{d}, \tag{3.8}$$

where:

- $a_{j,m}$ and $b_{j,m}$ are the linear regression parameters ($a_{j,m}$ represents the slope of the linear regression and $b_{j,m}$ represents the y-intercept of the linear regression);
- f_0 is the minimum intensity level considered in the linear regression;
- d is the number of intensities which are greater than or equal to f_0;
- $\mu_j[m]$ represents the average intensity of the normalized mass spectra of class c_j at mass-to-charge ratio m;
- $\sigma_j[m]$ represents the standard deviation of intensities at mass-to-charge ratio m for class c_j.

There are many small intensities and few large intensities in a mass spectrum. To compensate the imbalance of intensity levels in a mass spectrum and to avoid a biased regression model, only the intensity values greater than or equal to a certain level f_0 (for example 5%

of the base peak intensity) are considered in the summations of Equation (3.7) and Equation (3.8).

3.3.2 Normality Assessment

The intensities at mass-to-charge ratio m of the predefined class c_j are modeled to follow a Gaussian distribution. Normality test is performed before further statistical analysis of the data. The hypothesis that a Gaussian distribution models the data is rejected only if there is a strong evidence to the contrary.

An informal graphical approach to test normality is to compare a histogram of the data to a normal probability curve. The actual distribution of the histogram should be bell-shaped and resemble the normal distribution. A more formal graphical tool is the normal probability plot, a quantile-quantile plot [98] against the standard normal distribution. There are normality tests in statistics used to determine whether a dataset is well-modeled by a normal distribution. The Kolmogorov-Smirnov test [99] compares the cumulative distribution of the data with the expected cumulative normal distribution, and bases its p-value on the largest discrepancy. It turns out, however, that it is too simple, and does not adequately discriminate whether or not the data were sampled from a normal distribution. Other common normality tests include Anderson-Darling test [100], Jarque-Bera test [101], Shapiro-Wilk test [102], etc. In a thorough review of many available normality tests, D'Agostino [103] concluded that the most desirable procedure for testing hypothesis of normality is the D'Agostino-Pearson test [104] which assesses the normality using symmetry and kurtosis measures. The D'Agostino-Pearson test [105] is used in this study to determine whether a set of intensities are well-modeled by a Gaussian distribution.

D'Agostino-Pearson test

The D'Agostino-Pearson test assesses the normality using symmetry and kurtosis measures. Given n samples $\{z_i \mid i = 1, 2, \ldots, n\}$ and the mean μ, the third moment about the mean provides a measure of symmetry of the samples and is defined as:

$$k_3 = \frac{n \sum_{i=1}^{n} (z_i - \mu)^3}{(n-1)(n-2)}. \tag{3.9}$$

Because taking large or small numbers to their third or fourth powers can lead to serious rounding errors, computer algorithms may use a machine formula for k_3 as [105]:

$$k_3 = \frac{n \sum_{i=1}^{n} z_i^3 - 3 \sum_{i=1}^{n} z_i \sum_{i=1}^{n} z_i^2 + 2(\sum_{i=1}^{n} z_i)^3/n}{(n-1)(n-2)}. \tag{3.10}$$

The following normalized form is more commonly used as a measure of symmetry of the samples:

$$g_1 = \frac{k_3}{\sigma^3}, \tag{3.11}$$

where σ is the standard deviation.

A value for g_1 near 0 indicates that the samples come from a distribution symmetrically around the mean, one in which the mean and the median are identical and the frequency polygon to the left of the mean is a mirror image of the frequency polygon to the right of the mean, as shown in Figure 3.3 (a). A value for g_1 significantly less than 0 indicates that the samples come from a distribution that is skewed to the left, one in which the mean is less than the median, as shown in Figure 3.3 (b). A value for g_1 significantly larger than 0 indicates that the samples come from a distribution that is skewed to the right, one in which

the mean is larger than the median, as shown in Figure 3.3 (c).

The fourth power of the deviations from the mean provides a measure called kurtosis:

$$k_4 = \frac{\sum_{i=1}^{n}(z_i - \mu)^4 n(n+1)(n-1) - 3[\sum_{i=1}^{n}(z_i - \mu)^2]^2}{(n-2)(n-3)}. \tag{3.12}$$

The machine formula for k_4 is [105]:

$$k_4 = \frac{k_4'}{n(n-1)(n-2)(n-3)}, \tag{3.13}$$

where

$$\begin{aligned} k_4' = {} & (n^3 + n^2)\sum_{i=1}^{n} z_i^4 - 4(n^2+n)\sum_{i=1}^{n} z_i^3 \sum_{i=1}^{n} z_i - 3(n^2-n)(\sum_{i=1}^{n} z_i^2)^2 \\ & + 12n\sum_{i=1}^{n} z_i^2 (\sum_{i=1}^{n} z_i)^2 - 6(\sum_{i=1}^{n} z_i)^4. \end{aligned} \tag{3.14}$$

The following normalized form is more commonly used as a measure of kurtosis of the samples:

$$g_2 = \frac{k_4}{\sigma^4}. \tag{3.15}$$

A value for g_2 near 0 indicates that the samples come from a mesokurtic distribution, as shown in Figure 3.4 (a). A value for g_2 significantly less than 0 indicates that the samples come from a platykurtic distribution, as shown in Figure 3.4 (b). A value for g_2 significantly larger than 0 indicates that the samples come from a leptokurtic distribution, as shown in Figure 3.4 (c).

The sample symmetry g_1 and kurtosis g_2 are both asymptotically normal. However, the

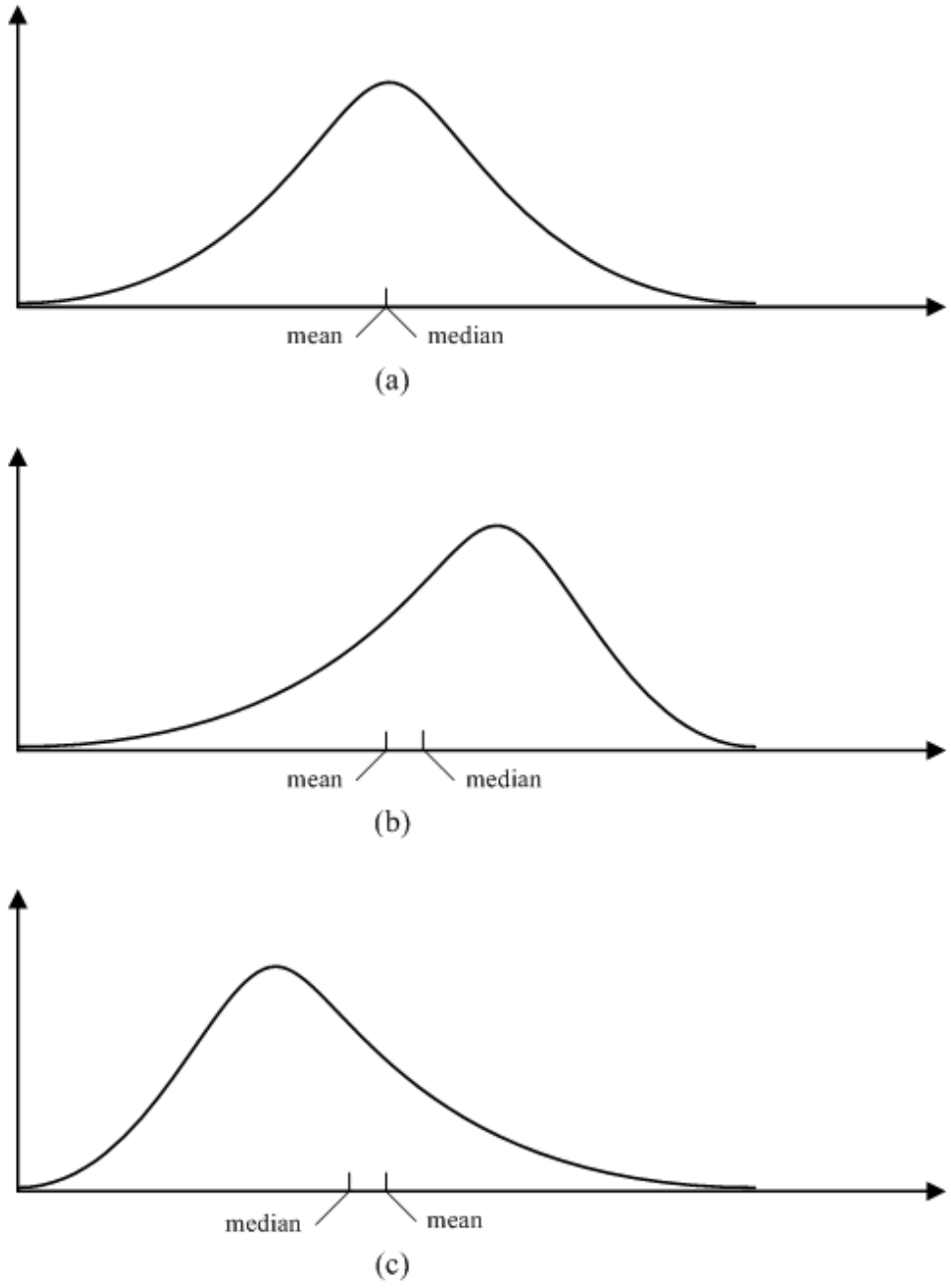

Figure 3.3: (a) Symmetrical distribution in which the mean and the median are identical. (b) Skewed to the left distribution in which the mean is less than the median. (c) Skewed to the right distribution in which the mean is larger than the median.

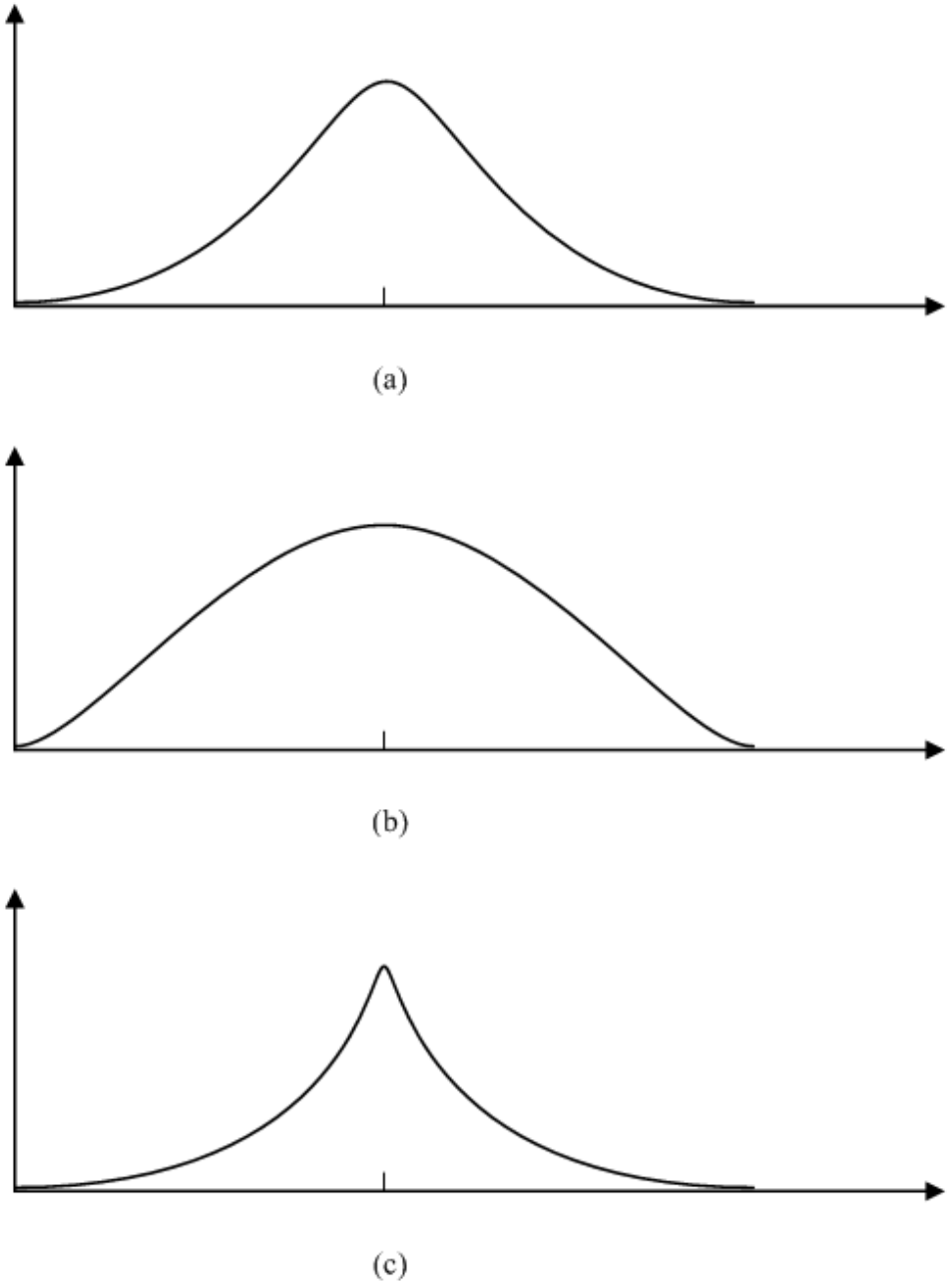

Figure 3.4: (a) Mesokurtic distribution. (b) Platykurtic distribution. (c) Leptokurtic distribution.

rate of their convergence to the distribution limit is frustratingly slow, especially for g_2. In order to remedy this situation, the D'Agostino-Pearson test uses the transformed symmetry Z_{g1} and kurtosis Z_{g2} which make distributions of g_1 and g_2 as close to standard normal as possible. D'Agostino [106] suggested transforming the sample symmetry g_1:

$$A = \frac{(n-2)g_1}{\sqrt{n(n-1)}}\sqrt{\frac{(n+1)(n+3)}{6(n-2)}}, \tag{3.16}$$

$$B = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}, \tag{3.17}$$

$$C = \sqrt{2(B-1)} - 1, \tag{3.18}$$

$$D = \sqrt{C}, \tag{3.19}$$

$$E = \frac{1}{\sqrt{\ln D}}, \tag{3.20}$$

$$F = \frac{A\sqrt{C-1}}{\sqrt{2}}, \tag{3.21}$$

$$Z_{g1} = E\ln(F + \sqrt{F^2+1}). \tag{3.22}$$

Z_{g1} is the statistic testing the null hypothesis of distribution symmetry.

Similarly, Anscombe and Glynn [107] suggested transforming the sample kurtosis g_2:

$$G = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}, \tag{3.23}$$

$$H = \frac{(n-2)(n-3)|g_2|}{(n+1)(n-1)\sqrt{G}}, \tag{3.24}$$

$$J = \frac{6(n^2-5n+2)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}, \tag{3.25}$$

$$K' = 6 + \frac{8}{J}\left(\frac{2}{J} + \sqrt{1 + \frac{4}{J^2}}\right), \tag{3.26}$$

$$L = \frac{1 - \frac{2}{K'}}{1 + H\sqrt{\frac{2}{K'-4}}}, \tag{3.27}$$

$$Z_{g2} = \frac{1 - \frac{2}{9K'} - \sqrt[3]{L}}{\sqrt{\frac{2}{9K'}}}. \tag{3.28}$$

Z_{g2} is the statistic testing the null hypothesis of distribution mesokurtosis.

The null hypothesis of distribution normality is tested using the statistic

$$K = Z_{g1}^2 + Z_{g2}^2. \tag{3.29}$$

K is a combination of statistics Z_{g1} and Z_{g2}. This omnibus test is able to detect deviations from normality due to either symmetry or kurtosis. K is approximately χ^2-distributed with 2 degrees of freedom under the null hypothesis of distribution normality, as the χ^2-distribution with ν degrees of freedom is the distribution of a sum of the squares of ν independent standard normal random variables. The cumulative distribution function of the χ^2-distribution is:

$$F_\nu(K) = \frac{\gamma(\frac{\nu}{2}, \frac{K}{2})}{\Gamma(\frac{\nu}{2})} = P(\frac{\nu}{2}, \frac{K}{2}), \tag{3.30}$$

where:

- $\gamma(\frac{\nu}{2}, \frac{K}{2})$ is an incomplete gamma function;
- $\Gamma(\frac{\nu}{2})$ is a gamma function;
- $P(\frac{\nu}{2}, \frac{K}{2})$ is a regularized gamma function.

In a special case of $\nu = 2$, this cumulative distribution function has a simple form:

$$F_2(K) = 1 - e^{-\frac{K}{2}}. \tag{3.31}$$

In statistical significance testing, the p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. As the cumulative distribution function of the χ^2-distribution for the appropriate degrees of freedom gives the probability of having obtained a value less extreme than this point, subtracting the cumulative distribution function value from 1 gives the p-value:

$$p\text{-}value(K) = 1 - F_2(K) = e^{-\frac{K}{2}}. \tag{3.32}$$

A p-value smaller than a significance level (denoted by α) is regarded as statistically significant. The lower the p-value, the less likely the result is if the null hypothesis is true, and consequently the more significant the result is, in the sense of statistical significance. The lower the significance level, the stronger the evidence required. Often used significance levels are 10% (0.1), 5% (0.05), 1% (0.01), and 0.1% (0.001) in many applications. A 0.1% (0.001) level of statistical significance implies there is only one chance in a thousand the result could have happened by coincidence. The null hypothesis of distribution normality cannot be rejected if the p-value is larger than the significance level α (for example 10%, 5%, 1%, or 0.1%). The alternative hypothesis (rejecting the null hypothesis) is accepted if the p-value is less than or equal to the significance level α (10%, 5%, 1%, or 0.1%) corresponding to a α (10%, 5%, 1%, or 0.1% respectively) chance of rejecting the null hypothesis when it is true (type I error).

3.4 Smoothing Model

The mean-centered intra-class variability model of class c_j at mass-to-charge ratio m:

$$\bar{\mathcal{N}}_{j,m}[f] = \int_{f-\frac{1}{2}}^{f+\frac{1}{2}} \frac{1}{\sqrt{2\pi}\,\sigma_j[m]} \exp(\frac{-x^2}{2\sigma_j^2[m]})dx \tag{3.33}$$

is used as a smoothing function to do a weighted moving average smoothing of the model expressing domain intensity characteristics at mass-to-charge ratio m. The smoothing width is from $-\infty$ to ∞. The smoothing is defined as:

$$(P_m * \bar{\mathcal{N}}_{j,m})[f] = \sum_{t=-\infty}^{\infty} P_m[f-t]\bar{\mathcal{N}}_{j,m}[t], \tag{3.34}$$

where:

- $P_m[f]$ represents the intensity probability of f at mass-to-charge ratio m;
- $\bar{\mathcal{N}}_{j,m}[f]$ represents the mean-centered intra-class variability model of class c_j at mass-to-charge ratio m;
- $P_m * \bar{\mathcal{N}}_{j,m}$ is the convolution of P_m and $\bar{\bar{\mathcal{N}}}_{j,m}$;
- f is the intensity;
- t is the intensity offset.

The weighted moving average is the integral of the overlapping region of $\bar{\mathcal{N}}_{j,m}[f]$ as it is shifted over $P_m[f]$. In the smoothing, each value of the domain characteristics model

$P_m[f]$ is weighted averaged with its neighbors:

$$\begin{aligned}(P_m * \bar{\mathcal{N}}_{j,m})[f] = \ldots &+ P_m[f+2]\bar{\mathcal{N}}_{j,m}[-2] + P_m[f+1]\bar{\mathcal{N}}_{j,m}[-1] \\ &+ P_m[f]\bar{\mathcal{N}}_{j,m}[0] + P_m[f-1]\bar{\mathcal{N}}_{j,m}[1] \\ &+ P_m[f-2]\bar{\mathcal{N}}_{j,m}[2] + \ldots .\end{aligned} \tag{3.35}$$

In the weighted average, $P_m[f]$ has the largest significance (largest weighting factor). The neighbors of $P_m[f]$, such as $P_m[f+1]$, $P_m[f-1]$, etc., have smaller significance than $P_m[f]$. The weighting factors determined by $\mathcal{N}_{j,m}[f]$ are symmetrical around $P_m[f]$. Nearer neighbors of $P_m[f]$ have larger weighting factors and further neighbors have smaller weighting factors.

3.5 Probability-Based Spectrum Similarity Measure

The probability-based spectrum similarity (PSS) between the query mass spectrum $\bar{x}_q$ and the labeled mass spectrum $\bar{x}_i$ (whose class label is y_i) is:

$$PSS[\bar{x}_q, \bar{x}_i] = \sum_{m=m_{min}}^{m_{max}} w_m \log\left(\frac{\bar{\mathcal{N}}_{y_i,m}[\bar{x}_q[m] - \bar{x}_i[m]]}{(P_m * \bar{\mathcal{N}}_{y_i,m})[\bar{x}_q[m]]}\right), \tag{3.36}$$

where:

- $PSS[\bar{x}_q, \bar{x}_i]$ is the probability-based spectrum similarity between base-peak normalized mass spectra $\bar{x}_q$ and $\bar{x}_i$;
- m_{min} is the minimum m/z of the query mass spectrum and the labeled mass spectra, and m_{max} is the maximum m/z of the query mass spectrum and the labeled mass spectra;

- w_m is the weight factor of the mass-to-charge ratio m and is either 1 or 0 as discussed in more detail later;
- $\bar{x}_q[m]$ is the base-peak normalized intensity of the query mass spectrum x_q at mass-to-charge ratio m;
- $\bar{x}_i[m]$ is the base-peak normalized intensity of the labeled mass spectrum x_i at mass-to-charge ratio m;
- $\bar{\mathcal{N}}_{y_i,m}[\bar{x}_q[m] - \bar{x}_i[m]]$ represents the mean-centered intra-class variability model for class y_i at mass-to-charge ratio m;
- $P_m[\bar{x}_q[m]]$ represents the domain characteristics model of labeled data at mass-to-charge ratio m;
- $P_m * \bar{\mathcal{N}}_{y_i,m}$ represents the smoothing model for class y_i at mass-to-charge ratio m.

The larger the PSS value, the larger probability that two spectra are similar and of the same class.

The denominator of the probability-based spectrum similarity is the domain characteristics of the labeled data at the query mass spectrum intensity level smoothed by the mean-centered intra-class variability of class y_i. As the domain intensity probability at the query mass spectrum intensity level increases (the intensity occurring more frequently in the domain), the probability-based spectrum similarity between two spectra decreases (because other classes are increasingly likely).

The numerator of the probability-based spectrum similarity is the mean-centered intra-class variability of class y_i at mass-to-charge ratio m used as a mapping function. In the

numerator, the normalized intensity difference between the query spectrum x_q and the labeled spectrum x_i at mass-to-charge ratio m is mapped to the measure of the similarity between the two spectra at mass-to-charge ratio m by the mean-centered intra-class variability model as shown in Figure 3.5. When two intensities are nearly the same, the intensity difference as a distance is near zero, and the numerator has the largest value. As the difference between the two intensities increases, the numerator decreases, therefore the probability-based spectrum similarity decreases.

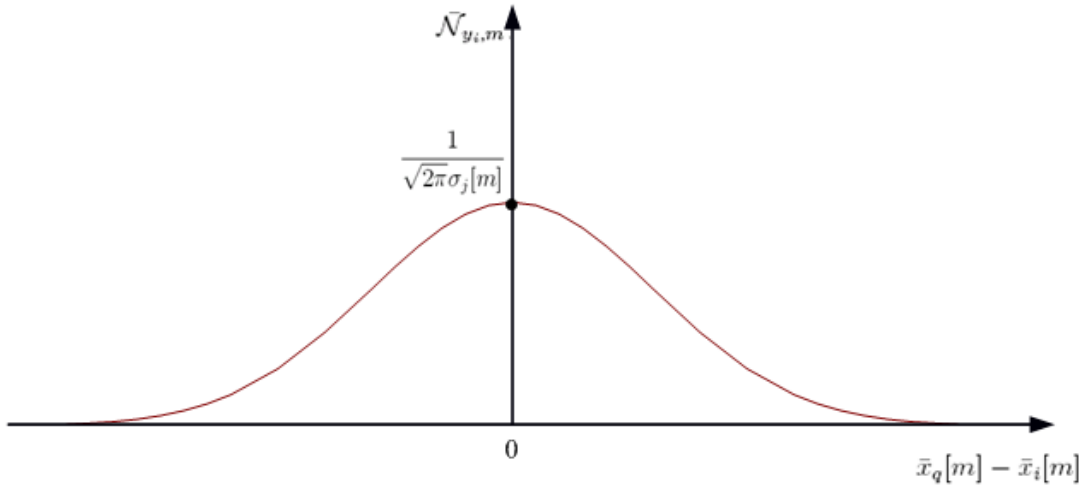

Figure 3.5: The normalized intensity difference between the query spectrum x_q and the labeled spectrum x_i is used as an offset in the mean-centered intra-class variability distribution to measure the similarity between two spectra.

In Equation (3.36), the logarithm of the probability at each mass-to-charge ratio m is summed instead of probability multiplication, because multiplication of small probabilities leads to mathematical underflow, but addition of logarithm probabilities does not. The summation includes all m/z from m_{min} to m_{max}. Each m/z has a weight factor w_m. For a mass-to-charge ratio m, if all the labeled spectra have the same intensity, this mass-to-charge ratio m does not contain any information to discriminate classes. In this case, the

weight factor w_m of the mass-to-charge ratio m is set to be zero. And the weight factors of m/z whose intensity distributions do not pass the D'Agostino-Pearson normality test in the intra-class variability model are set to be zero. All other m/z have equal significance with $w_m = 1$. This probability-based spectrum similarity measure considers all m/z that contain discriminating information to avoid information loss.

Chapter 4

Experimental Results for the MSN-PSSM Algorithm

Experimental results demonstrate the effectiveness and robustness of the new MSN-PSSM algorithm. MSN-PSSM outperforms popular classification techniques for classification of mass spectra, such as PCA with DFA, SIMCA, and decision tree learning.

4.1 Datasets

The first test dataset for performance evaluation was acquired with a quadrupole GC×GC-MS instrument by Dr. Edward B. Ledford at Zoex Corp. [108]. Figure 4.1 illustrates the GC×GC-MS image of the mixture of compounds containing paraffins, isoparaffins, aromatics, naphthenes, and olefins (PIANO) with color annotated blobs considered for chemical classification. Paraffins, isoparaffins, aromatics, naphthenes, and olefins are important

categories of chemicals.

- Paraffins are straight-chain (linear) alkane hydrocarbons with the general formula C_nH_{2n+2}, such as pentane, hexane, etc.
- Isoparaffins are branched-chain alkane hydrocarbons with the general formula C_nH_{2n+2}, such as isopentane, 3-methylpentane, etc.
- Aromatics contain one or more benzene rings, such as benzene, toluene, etc.
- Naphthenes (cycloalkanes) are types of alkane hydrocarbons which have one or more rings of carbon atoms, such as indane, cyclopentane, etc.
- Olefins (alkenes) are unsaturated chemical compounds containing at least one carbon-to-carbon double bond, such as ethene, isobutene, etc.

These five categories of chemicals are interlaced in the image, as a result, GC×GC with only retention times of two columns may not be sufficient to discriminate these chemicals. GC×GC-MS combines two techniques (GC×GC and MS) providing enhanced capability for chemical identification. It potentially can discriminate chemicals better than GC×GC. In Figure 4.1, eight blue blobs are paraffins, thirteen green blobs are isoparaffins, thirty-three violet blobs are aromatics, twenty red blobs are naphthenes, and eleven yellow blobs are olefins. Each spectrum is from m/z 45 to 221, and binned to 1 mass intervals from -0.5 to +0.5 of each integer mass. Discrimination of these PIANO chemicals is a challenging unbalanced multi-class supervised classification task to demonstrate structure discrimination for chemical compounds.

The second test dataset for performance evaluation was acquired from the NIST/EPA/NIH Mass Spectral Library 2005 (NIST05) [109]. NIST05 provides a collection of known

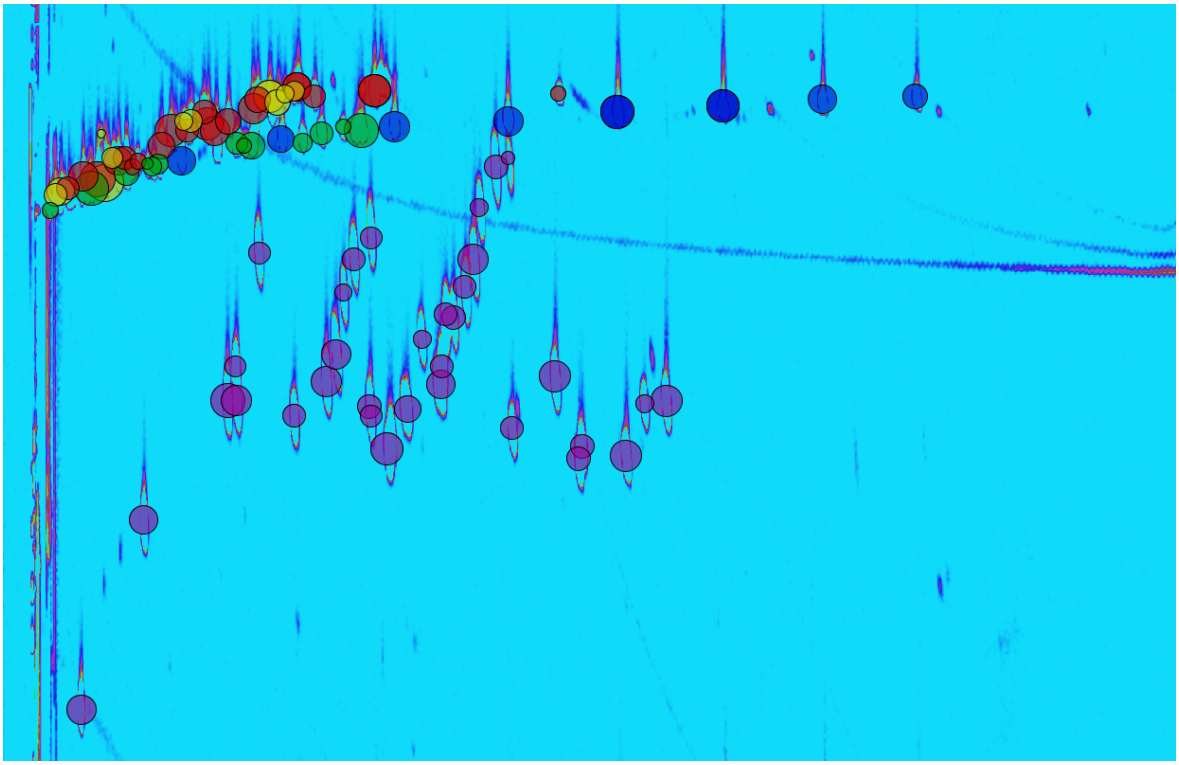

Figure 4.1: GC×GC-MS image of the PIANO mixture. Eight blue blobs are paraffins, thirteen green blobs are isoparaffins, thirty-three violet blobs are aromatics, twenty red blobs are naphthenes, and eleven yellow blobs are olefins.

chemical compounds with their mass spectra. This library is the product of a multi-year, comprehensive evaluation and expansion of the world's most widely used mass spectral reference library. It contains 163198 mass spectra with their compound names, formulas, chemical structures, and other information. Each mass spectrum of this library is critically examined by experienced mass spectrometrists. The second test dataset includes five categories of chemicals: forty mass spectra of parafins, thirty-five isoparaffins, thirty-nine aromatics, twenty-six naphthenes, and twenty-eight olefins (PIANO). The composition of each category is based on PIANO calibration standards from Air Liquide American Specialty Gases LLC [110] and definitions. Each mass spectrum is from m/z 1 to 619, and binned to 1 mass intervals from -0.5 to +0.5 of each integer mass. Discrimination of these PIANO chemicals is a challenging unbalanced multi-class supervised classification task to demonstrate structure discrimination for chemical compounds.

The third test dataset for performance evaluation was acquired by Dr. John C. Vickerman's group with a BioToF-SIMS instrument at the Surface Analysis Research Centre, University of Manchester. This test dataset includes ToF-SIMS spectra of bacterial samples which are major causal agents of urinary tract infection (UTI). UTI is a serious health problem affecting millions of people each year [111]. There is a growing need to identify the causal agent prior to treatment. This UTI dataset has samples for sixteen strains (categories) of UTI bacteria. These sixteen strains are five strains of *Escherichia coli* (Eco), one strain of *Klebsiella oxytoca* (Kox), three strains of *Klebsiella pneumoniae* (Kpn), two strains of *Citrobacter freundii* (Cfr), four strains of *Enterococcus spp* (Esp), and one strain of *Proteus mirabilis* (Pmi). Each strain has three biological replicates from three fresh areas of each bacterial sample. Three ToF-SIMS spectra are generated for each biological replicate to make three machine replicates. Thus, in total there are nine ToF-SIMS spectra for each of sixteen strains of UTI bacteria. These strains were previously identified by con-

ventional biochemical tests. Bacterial sample growth, ToF-SIMS instrumentation, and data acquisition parameters are described in detail by Fletcher [38]. Each ToF-SIMS spectrum is from m/z 1 to 1000, and binned to 1 mass intervals from -0.5 to +0.5 of each integer mass. The mass spectra are highly similar and enormously complex, having many peaks with varying intensities over mass range 1 to 1000. Many common peaks make visual inspection and manual identification of spectra an impossible task. Hence, it is necessary to develop automatic techniques to analyze these complex data. Discrimination of these UTI bacteria is a challenging multi-class supervised classification task to demonstrate strain-level discrimination for the subtly different bacterial samples.

4.2 Pre-processing

The spectra of the UTI dataset are dominated by Na^+ (m/z=23) and K^+ (m/z=39) ions. Because this salt contamination is apparent and peaks in the low mass region have little discrimination ability, m/z from 1 to 50 are pruned from the UTI dataset spectra. Without any other pruning, each mass spectrum of the three datasets is normalized to the most intense peak (the base peak) of the spectrum.

4.3 Performance Evaluation

This study uses cross-validation, a popular accuracy estimation technique [112], to assess prediction accuracy. Given a classification algorithm and a dataset, k-fold cross-validation splits the data into k approximately equally sized partitions, or folds. The classification

algorithm is executed k times. Each time, a classifier is trained on k-1 folds and the generated hypothesis is tested on the unseen fold, which serves as a test set. The estimated accuracy is computed as the average accuracy over the k test sets.

Leave-one-out cross-validation is almost unbiased [113] and is commonly considered the preferred method for splitting the dataset. Leave-one-out cross-validation, which is a commonly used technique in chemometrics, is adopted in this study. Overall supervised classification accuracy with leave-one-out cross-validation is defined as:

$$\text{Accuracy} = \frac{\text{\# of spectra classified correctly}}{\text{\# of spectra in the dataset}}. \tag{4.1}$$

Overall supervised classification accuracy with leave-one-out cross-validation is used to quantitatively measure the performance of supervised classification algorithms.

4.4 Significance Assessment

Given a supervised classification algorithm, this study uses the binomial test of significance [114] and Fleiss kappa statistic [115, 116] to quantitatively measure the significance of the classification algorithm's performance. Given two supervised classification algorithms, this study uses paired t-test [117] to quantitatively measure the significance of the performance difference between the two classification algorithms, that is, how significant the performance difference is.

4.4.1 Binomial Test of Significance

Given n data and p predefined classes, random guessing would be able to give a $p_0 = \frac{1}{p}$ probability of correct prediction for each datum without a priori knowledge of class distribution. For a classification algorithm which correctly classifies r $(r \leq n)$ of the n data, the binomial test of significance tests that how likely the algorithm's achievement is relative to random guessing.

The first step of the binomial test of significance is stating the hypotheses. The null hypothesis states that the classification algorithm does not have any ability to tell the difference between the p predefined classes, that is, the classification algorithm classifies each datum purely based on random guessing (without a priori knowledge of class distribution):

$$H_0 : p_0 = \frac{1}{p}, \tag{4.2}$$

where p_0 is the probability of correct prediction for each datum without a priori knowledge of the class distribution. The alternative hypothesis is that the classification algorithm does have the ability to tell the difference between the p predefined classes better than random guessing:

$$H_a : p_0 > \frac{1}{p}, \tag{4.3}$$

corresponding to a one-tailed binomial test which only examines differences in only one of two possible directions.

The second step of the binomial test of significance is calculating the test statistic. The probability of getting a correct classification result for a datum by random guessing (without a priori knowledge of class distribution) is $p_0 = \frac{1}{p}$, and the probability of getting a

wrong classification result for a datum by a random guessing is $p_1 = 1 - \frac{1}{p}$. The probability of getting r correct classification results over n data is:

$$P(r) = C_n^r \, p_0^r \, p_1^{n-r} = \frac{n!}{r!(n-r)!} \, p_0^r \, (1-p_0)^{n-r}. \tag{4.4}$$

The probability of getting r or more than r correct classification results over n data is:

$$p\text{-}value = P(r) + P(r+1) + \ldots + P(n). \tag{4.5}$$

When the sample size n is sufficiency large and p_0 is not too close to 0 or 1, according to the central limit theorem [118], the normal distribution with np_0 as the mean and np_0p_1 as the variance is a good approximation to the binomial distribution. How large n needs to be depends on the value of p_0. If p_0 is near 0.5, the approximation can be good for n much less than 20 [105]. However, it is better to be conservative and limit the use of the normal distribution as an approximation to the binomial distribution when $np_0p_1 \geq 5$ [35]. In the normal approximation of the binomial test, a normal curve Z-test is used as an approximation of the binomial test, using this formula:

$$Z = \frac{r - n\,p_0}{\sqrt{n\,p_0\,p_1}} = \frac{r - \frac{n}{p}}{\sqrt{\frac{n}{p}(1 - \frac{1}{p})}}, \tag{4.6}$$

where np_0 is the mean of the distribution and np_0p_1 is the variance of the distribution.

The third step of the binomial test of significance is setting a significance level α. The lower the significance level, the stronger the evidence required. Often used significance levels are 10% (0.1), 5% (0.05), 1% (0.01), and 0.1% (0.001). A 0.1% (0.001) level of statistical significance implies there is only one chance in a thousand that the classification

algorithm correctly classifies r of the n data by coincidence.

The fourth step of the binomial test of significance is making a decision about the null hypothesis and stating a conclusion. The alternative hypothesis (the classification algorithm does have the ability to tell the difference between the p predefined classes better than random guessing) is accepted if the p-value is less than or equal to the significance level α, corresponding to a α chance of rejecting the null hypothesis when it is true. The null hypothesis of random guessing cannot be rejected if the p-value is larger than the significance level α.

In the normal approximation of the binomial test, a critical value (a value that a test statistic must exceed in order for the null hypothesis to be rejected) can be determined based on:

$$Z_{\alpha(1)} = t_{\alpha,\infty}, \tag{4.7}$$

where:

- $Z_{\alpha(1)}$ represents the critical value of the one-tailed Z-test with significance level α;
- $t_{\alpha,\infty}$ represents the critical value of the t-distribution with significance level α and degree of freedom ∞.

The critical value of the one-tailed Z-test with 0.1 significance level is $Z_{0.1(1)} = 1.2816$ based on the critical value table of the t-distribution [105]. The critical value with 0.05 significance level is $Z_{0.05(1)} = 1.645$. The critical value with 0.01 significance level is $Z_{0.01(1)} = 2.326$. The critical value with 0.001 significance level is $Z_{0.001(1)} = 3.090$. The alternative hypothesis (the classification algorithm does have the ability to tell the difference between the p predefined classes better than random guessing) is accepted if Z

is greater than or equal to $Z_{\alpha(1)}$. The null hypothesis of random guessing cannot be rejected if Z is less than $Z_{\alpha(1)}$.

4.4.2 Fleiss Kappa Statistic

The kappa statistic is a widely used method to measure the reliability of agreement between raters or classifiers [119, 120]. It is a chance-corrected statistical measure which calculates the reliability of agreement in classification over that which would be expected by chance. The kappa statistic was first proposed by Cohen [121]. Some extensions were developed by others, including Cohen [122], Everitt [123], and Fleiss [115].

This study uses Fleiss kappa statistic [116] to measure the agreement between the predicted class labels from a supervised classification algorithm and the true class labels of data. Fleiss kappa statistic can be interpreted as expressing the extent to which the observed agreement between the predicted class labels from the classification algorithm and the true class labels of data exceeds what would be expected if the classification algorithm made its decision randomly.

Given n data $\{x_i \mid i = 1, 2, \ldots, n\}$ and p predefined mutually exclusive and exhaustive classes $\{c_j \mid j = 1, 2, \ldots, p\}$, the Fleiss kappa statistic between the predicted class labels and the true class labels is [116]:

$$kappa = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}}, \tag{4.8}$$

where:

- $kappa$ is the Fleiss kappa statistic;

- the factor $1 - \overline{P_e}$ measures the reliability of agreement attainable above chance;
- the factor $\overline{P} - \overline{P_e}$ measures the reliability of agreement actually achieved above chance.

To define $\overline{P}$ and $\overline{P_e}$, let f_{ij} represent the frequency of datum x_i having c_j as its label (both the predicted class label and the true class label). For example, given a case of data x_1 and p predefined classes $\{c_j \mid j = 1, 2, \ldots, p\}$, if x_1's predicted class label (from a supervised classification algorithm) is c_1 and its true class label is c_2, then $f_{11} = f_{12} = 1$ and $f_{13} = \ldots = f_{1p} = 0$. For x_i, $\sum_{j=1}^{p} f_{ij} = 2$, because each case of data has two class labels (one predicted class label and one true class label). For all n data, the proportion of label c_j in the predicted class labels and the true class labels ($2n$ labels) is:

$$Q_j = \frac{1}{2n} \sum_{i=1}^{n} f_{ij}. \tag{4.9}$$

For x_i, the extent to which its predicted class label agrees with its true class label is defined as:

$$\begin{aligned} P_i &= \frac{1}{2} \sum_{j=1}^{p} f_{ij}(f_{ij} - 1) \\ &= \frac{1}{2} \sum_{j=1}^{p} (f_{ij}^2 - f_{ij}) \\ &= \frac{1}{2} \left(\sum_{j=1}^{p} f_{ij}^2 - 2 \right). \end{aligned} \tag{4.10}$$

For all n data, the mean of the P_i is:

$$\begin{aligned}\overline{P} &= \frac{1}{n}\sum_{i=1}^{n} P_i \\ &= \frac{1}{2n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} f_{ij}^2 - 2\right) \\ &= \frac{1}{2n}\left(\sum_{i=1}^{n}\sum_{j=1}^{p} f_{ij}^2 - 2n\right).\end{aligned} \tag{4.11}$$

And $\overline{P_e}$ is defined as:

$$\overline{P_e} = \sum_{j=1}^{p} Q_j^2. \tag{4.12}$$

The kappa result ranges from -1 to 1. A negative kappa value occurs when agreement is weaker than expected by chance. Higher kappa values mean stronger agreement. A kappa value of 1 means perfect agreement. Interpretation of the kappa values is based on Landis's categories, shown in Table 4.1.

Table 4.1: Interpretation of kappa values [124].

Kappa values	**Interpretation**
< 0	Poor agreement
0.00-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1.00	Almost perfect agreement

4.4.3 Paired t-test

Given two paired sets of measured values, the paired t-test determines whether they differ from each other in a significant way under the assumptions that the differences of the paired values are independent and normally distributed. This study uses k-fold cross-validated paired t-test [125] to quantitatively measure the significance of the performance difference between two supervised classification algorithms.

The first step of the k-fold cross-validated paired t-test is stating the hypothesis. In the k-fold cross-validation, data are randomly split into k approximately equally sized folds, and classification algorithms are executed k times. Each time, classifiers are trained on $k-1$ folds and tested on the unseen fold, which serves as a test set. The classification error rate of a classification algorithm is the number of test data classified wrongly divided by the number of test data. Given n data and two classification algorithms, the k-fold cross-validation generates two sets of classification error rates $\{e_i \mid i = 1, 2, \ldots, k\}$ and $\{f_i \mid i = 1, 2, \ldots, k\}$, one for each classification algorithm. The difference in error rate on fold i is:

$$g_i = e_i - f_i. \tag{4.13}$$

According to the central limit theorem [118], when k is sufficiency large, both $\{e_i \mid i = 1, 2, \ldots, k\}$ and $\{f_i \mid i = 1, 2, \ldots, k\}$ are approximately normally distributed. Therefore the difference $\{g_i \mid i = 1, 2, \ldots, k\}$ also is approximately normally distributed. The null hypothesis is that the two supervised classification algorithms have the same classification performance (the same error rate), that is, $\{g_i \mid i = 1, 2, \ldots, k\}$ is from a distribution with zero mean:

$$H_0 : \mu_g = 0, \tag{4.14}$$

where μ_g is the mean of the distribution. The alternative hypothesis is that the two supervised classification algorithms have different classification performance (different error rate), that is, $\{g_i \mid i = 1, 2, \ldots, k\}$ is from a distribution with non-zero mean:

$$H_a : \mu_g \neq 0, \tag{4.15}$$

corresponding to a two-tailed t-test which examines differences in both of the two possible directions.

The second step of the k-fold cross-validated paired t-test is calculating the test statistic. If random samples are drawn from a normal distribution, the means of these samples will conform to a normal distribution. The distribution of means from a non-normal distribution will tend toward normality as the size of samples increases [105]. Because $\{g_i \mid i = 1, 2, \ldots, k\}$ is approximately normal distributed, the mean of $\{g_i \mid i = 1, 2, \ldots, k\}$ is drawn from a normal distribution. Define the mean of $\{g_i \mid i = 1, 2, \ldots, k\}$ as:

$$\overline{g} = \frac{1}{k} \sum_{i=1}^{k} g_i, \tag{4.16}$$

and

$$\sigma_{\overline{g}} = \frac{\sigma}{\sqrt{k}} = \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (g_i - \overline{g})^2}, \tag{4.17}$$

where:

- $\sigma_{\overline{g}}$ is the estimated standard deviation of the distribution of means;
- σ is the standard deviation of $\{g_i \mid i = 1, 2, \ldots, k\}$:

$$\sigma = \sqrt{\frac{1}{k-1} \sum_{i=1}^{k} (g_i - \overline{g})^2}. \tag{4.18}$$

The t-distribution is a bell-shaped distribution similar to the normal distribution, but wider and shorter to reflect the greater variance introduced by using $\sigma_{\overline{g}}$ to approximate the true standard deviation. Define ν as the degree of freedom of the t-distribution. The t-distribution approaches the normal distribution as ν approaches infinity. Under the null hypothesis that $\mu_g = 0$, the following statistic is t-distributed with $\nu = k - 1$:

$$t = \frac{\overline{g} - \mu_g}{\sigma_{\overline{g}}} = \frac{\overline{g}}{\sigma_{\overline{g}}}. \tag{4.19}$$

The third step of the k-fold cross-validated paired t-test is setting a significance level. The smaller the significance level, the stronger the evidence required. Often used significance levels are 10% (0.1), 5% (0.05), 1% (0.01), and 0.1% (0.001). A 0.1% (0.001) level of statistical significance implies there is only one chance in a thousand the result could have happened by coincidence.

The fourth step of the k-fold cross-validated paired t-test is making a decision about the null hypothesis and stating a conclusion. Most statistical textbooks list the critical value table of the t-distribution. Table 4.2 lists a few selected two-tailed critical values $t_{\alpha(2),\nu}$ for t-distributions with ν degrees of freedom and significance level of $\alpha(2)$. The alternative hypothesis, the two supervised classification algorithms have different classification performance (different error rate), is accepted if t is greater than or equal to $t_{\alpha(2),k-1}$. The null hypothesis, the two supervised classification algorithms have the same classification performance (the same error rate), can not be rejected if t is less than $t_{\alpha(2),k-1}$.

Table 4.2: Selected critical values of the t-distribution.

ν	$\alpha(2) = 0.1$	$\alpha(2) = 0.05$	$\alpha(2) = 0.01$	$\alpha(2) = 0.001$
1	6.314	12.706	63.657	636.619
2	2.920	4.303	9.925	31.599
3	2.353	3.182	5.841	12.924
4	2.132	2.776	4.604	8.610
5	2.015	2.571	4.032	6.869
6	1.943	2.447	3.707	5.959
7	1.895	2.365	3.499	5.408
8	1.860	2.306	3.355	5.041
9	1.833	2.262	3.250	4.781
10	1.812	2.228	3.169	4.587
20	1.725	2.086	2.845	3.850
30	1.697	2.042	2.750	3.646
40	1.684	2.021	2.704	3.551
50	1.676	2.009	2.678	3.496
60	1.671	2.000	2.660	3.460
70	1.667	1.994	2.648	3.435
80	1.664	1.990	2.639	3.416
90	1.662	1.987	2.632	3.402
100	1.660	1.984	2.626	3.390
110	1.659	1.982	2.621	3.381
120	1.658	1.980	2.617	3.373
130	1.657	1.978	2.614	3.367
140	1.656	1.977	2.611	3.361
150	1.655	1.976	2.609	3.357
160	1.654	1.975	2.607	3.352
170	1.654	1.974	2.605	3.349
180	1.653	1.973	2.603	3.345
190	1.653	1.973	2.602	3.342
200	1.653	1.972	2.601	3.340
∞	1.6449	1.9600	2.5758	3.2905

4.5 Experimental Results

The five categories of compounds (paraffins, isoparaffins, aromatics, naphthenes, and olefins) of the PIANO datasets are considered as five classes with class labels Para, Isopara, Arom, Naph, and Olef. For the first PIANO dataset, Para has eight GC×GC-MS spectra, Isopara has thirteen spectra, Arom has thirty-three spectra, Naph has twenty spectra, and Olef has eleven spectra. For the second PIANO dataset, Para has forty mass spectra, Isopara has thirty-five spectra, Arom has thirty-nine spectra, Naph has twenty-six spectra, and Olef has twenty-eight spectra.

The sixteen stains of the UTI dataset are considered as sixteen classes with class labels Cfr1, Cfr2, Eco1, Eco2, Eco3, Eco4, Eco5, Esp1, Esp2, Esp3, Esp4, Kox, Kpn1, Kpn2, Kpn3, and Pmi. Each class has nine ToF-SIMS spectra.

These datasets are classified by four supervised classification algorithms: PCA with DFA, SIMCA, decision tree learning, and MSN-PSSM. PCA with DFA is implemented in Matlab (the MathWorks Inc.). PCA considers above 90% of the variance of the datasets. The principal components containing significantly small portions of the total variance capture noise variations, thus are not considered in PCA. SIMCA is implemented in Matlab. C4.5, designed by Quinlan [67], is employed to build classification trees and the variable selection measure in each node is gain ratio to avoid broad decision trees of small depth.

4.5.1 The First PIANO Dataset

Table 4.3 shows the overall classification accuracy and Fleiss kappa statistic of each supervised classification algorithm on the first PIANO dataset. The MSN-PSSM algorithm

outperforms the other three algorithms with the highest overall classification accuracy of 87.06%. SIMCA and decision tree learning have the same overall classification accuracy of 71.76%. PCA with DFA has 80.00% overall classification accuracy which is better than SIMCA and decision tree learning. The performance of these four classification algorithms would be achieved by random guessing (without a priori knowledge of class distribution) with less than 1.0×10^{-16} probability (greater than 99.9999% significance level) based on the binomial test of significance. The spectra's predicted class labels of MSN-PSSM have almost perfect agreement with the spectra's true class labels based on Table 4.1. The spectra's predicted class labels of the other three supervised classification algorithms hold only substantial agreement with the spectra's true class labels based on Table 4.1. The MSN-PSSM algorithm significantly outperforms the PCA with DFA algorithm at 86.55% significance level based on the paired t-test. MSN-PSSM significantly outperforms SIMCA at 99.94% significance level, and significantly outperforms decision tree learning at 99.76% significance level.

Table 4.4 shows the confusion matrix, the precision of each class, and the recall of each class for each supervised classification algorithm on the first PIANO dataset. Each row of the confusion matrix represents the data in the true class label and each column represents the data in a predicted class label. The diagonal elements show the number of correct

Table 4.3: Performance of classifiers on the first PIANO dataset. Boldface indicates the best performance.

Classifier	**Accuracy (%)**	**Kappa**
PCA with DFA	80.00	0.73
SIMCA	71.76	0.61
Decision trees	71.76	0.62
MSN-PSSM	**87.06**	**0.82**

classifications made for each class, and the off-diagonal elements show the errors made. Precision measures the accuracy that a specific class has been predicted. Recall measures the ability of a classification algorithm to select instances of a certain class from a data set. Table 4.4 shows MSN-PSSM can successfully discriminate between PIANO categories, and MSN-PSSM outperforms PCA with DFA, SIMCA, and decision tree learning.

Table 4.5 shows the classification results of the four algorithms on class Para of the first PIANO dataset. The MSN-PSSM algorithm and the PCA with DFA algorithm classify most spectra correctly. MSN-PSSM misclassifies two spectra and PCA with DFA misclassifies one spectrum. MSN-PSSM and PCA with DFA are able to successfully discriminate class Para from other classes. SIMCA and decision tree learning cannot discriminate class Para from other classes. SIMCA correctly classifies only two spectra and decision tree learning correctly classifies only three spectra. Most misclassified spectra are classified as class Isopara, which is consistent with the close structural similarity between class Para (straight-chain alkane) and class Isopara (branched-chain alkane).

Table 4.6 shows the classification results of the four algorithms on class Isopara of the first PIANO dataset. The MSN-PSSM algorithm classifies all spectra correctly. MSN-PSSM is able to perfectly discriminate class Isopara from other classes. PCA with DFA, SIMCA, and decision tree learning have similar performance. PCA with DFA misclassifies four spectra, SIMCA misclassifies four spectra, and decision tree learning misclassifies six spectra. Most misclassified spectra are classified as class Para, which is consistent with the close structural similarity between class Isopara and class Para.

Table 4.7 shows the classification results of the four algorithms on class Arom of the first PIANO dataset. The MSN-PSSM algorithm and the PCA with DFA algorithm classify all spectra correctly. MSN-PSSM and PCA with DFA are able to perfectly discrim-

Table 4.4: Confusion matrix, precision, and recall of each classification algorithm on the first PIANO dataset.

Classifier	**Confusion matrix**							
PCA with DFA		Para	Isopara	Arom	Naph	Olef	error	recall
	Para	7	1	0	0	0	1	0.88
	Isopara	4	9	0	0	0	4	0.69
	Arom	0	0	33	0	0	0	1.00
	Naph	0	2	0	11	7	9	0.55
	Olef	0	0	0	3	8	3	0.73
	error	4	3	0	3	7	17	
	precision	0.64	0.75	1.00	0.79	0.53		
SIMCA		Para	Isopara	Arom	Naph	Olef	error	recall
	Para	2	1	0	5	0	6	0.25
	Isopara	0	9	0	4	0	4	0.69
	Arom	0	0	32	1	0	1	0.97
	Naph	0	5	0	15	0	5	0.75
	Olef	0	7	0	1	3	8	0.27
	error	0	13	0	11	0	24	
	precision	1.00	0.41	1.00	0.58	1.00		
Decision trees		Para	Isopara	Arom	Naph	Olef	error	recall
	Para	3	5	0	0	0	5	0.38
	Isopara	4	7	0	2	0	6	0.54
	Arom	1	0	32	0	0	1	0.97
	Naph	0	2	0	15	3	5	0.75
	Olef	0	1	0	6	4	7	0.36
	error	5	8	0	8	3	24	
	precision	0.38	0.47	1.00	0.65	0.57		
MSN-PSSM		Para	Isopara	Arom	Naph	Olef	error	recall
	Para	6	2	0	0	0	2	0.75
	Isopara	0	13	0	0	0	0	1.00
	Arom	0	0	33	0	0	0	1.00
	Naph	0	1	1	18	0	2	0.90
	Olef	0	0	0	7	4	7	0.36
	error	0	3	1	7	0	11	
	precision	1.00	0.81	0.97	0.72	1.00		

Table 4.5: Classification results of the four algorithms on class Para of the first PIANO dataset.

ID	PCA with DFA	SIMCA	Decision trees	MSN-PSSM	True class label
1	Para	Para	Para	Para	Para
2	Para	Para	Para	Para	Para
3	Para	Naph	Para	Para	Para
4	Para	Naph	Isopara	Para	Para
5	Para	Naph	Isopara	Para	Para
6	Para	Naph	Isopara	Para	Para
7	Para	Naph	Isopara	Isopara	Para
8	Isopara	Isopara	Isopara	Isopara	Para
correct	7	2	3	6	

Table 4.6: Classification results of the four algorithms on class Isopara of the first PIANO dataset.

ID	PCA with DFA	SIMCA	Decision trees	MSN-PSSM	True class label
1	Para	Naph	Para	Isopara	Isopara
2	Isopara	Isopara	Isopara	Isopara	Isopara
3	Isopara	Isopara	Para	Isopara	Isopara
4	Isopara	Naph	Isopara	Isopara	Isopara
5	Para	Isopara	Isopara	Isopara	Isopara
6	Para	Isopara	Para	Isopara	Isopara
7	Isopara	Isopara	Naph	Isopara	Isopara
8	Isopara	Isopara	Para	Isopara	Isopara
9	Para	Isopara	Isopara	Isopara	Isopara
10	Isopara	Isopara	Isopara	Isopara	Isopara
11	Isopara	Naph	Isopara	Isopara	Isopara
12	Isopara	Naph	Naph	Isopara	Isopara
13	Isopara	Isopara	Isopara	Isopara	Isopara
correct	9	9	7	13	

inate class Arom from other classes. SIMCA and decision tree learning have the same performance of classifying most spectra correctly with only one misclassified spectrum. SIMCA and decision tree learning are able to successfully discriminate class Arom from other classes. The fact that class Arom is easy to discriminate from other classes is consistent with the unique benzene ring structure of class Arom.

Table 4.8 shows the classification results of the four algorithms on class Naph of the first PIANO dataset. The MSN-PSSM algorithm classifies most spectra correctly with only two misclassified spectra. MSN-PSSM is able to successfully discriminate class Naph from other classes. SIMCA and decision tree learning have the same performance of classifying most spectra correctly with five misclassified spectra. The PCA with DFA algorithm has the worst performance compared with MSN-PSSM, SIMCA, and decision tree learning. PCA with DFA correctly classifies only about half of the spectra (11 of the 20).

Table 4.9 shows the classification results of the four algorithms on class Olef of the first PIANO dataset. The PCA with DFA algorithm has the best performance of correctly classifying all but three spectra. MSN-PSSM and decision tree learning have the same performance of correctly classifying four spectra. SIMCA has the worst performance compared with PCA with DFA, MSN-PSSM, and decision tree learning. SIMCA correctly classifies only three spectra.

4.5.2 The Second PIANO Dataset

Table 4.10 shows the overall classification accuracy and Fleiss kappa statistic of each supervised classification algorithm on the second PIANO dataset. The MSN-PSSM algorithm outperforms the other three algorithms with the highest overall classification accu-

Table 4.7: Classification results of the four algorithms on class Arom of the first PIANO dataset.

ID	PCA with DFA	SIMCA	Decision trees	MSN-PSSM	True class label
1	Arom	Naph	Para	Arom	Arom
2	Arom	Arom	Arom	Arom	Arom
3	Arom	Arom	Arom	Arom	Arom
4	Arom	Arom	Arom	Arom	Arom
5	Arom	Arom	Arom	Arom	Arom
6	Arom	Arom	Arom	Arom	Arom
7	Arom	Arom	Arom	Arom	Arom
8	Arom	Arom	Arom	Arom	Arom
9	Arom	Arom	Arom	Arom	Arom
10	Arom	Arom	Arom	Arom	Arom
11	Arom	Arom	Arom	Arom	Arom
12	Arom	Arom	Arom	Arom	Arom
13	Arom	Arom	Arom	Arom	Arom
14	Arom	Arom	Arom	Arom	Arom
15	Arom	Arom	Arom	Arom	Arom
16	Arom	Arom	Arom	Arom	Arom
17	Arom	Arom	Arom	Arom	Arom
18	Arom	Arom	Arom	Arom	Arom
19	Arom	Arom	Arom	Arom	Arom
20	Arom	Arom	Arom	Arom	Arom
21	Arom	Arom	Arom	Arom	Arom
22	Arom	Arom	Arom	Arom	Arom
23	Arom	Arom	Arom	Arom	Arom
24	Arom	Arom	Arom	Arom	Arom
25	Arom	Arom	Arom	Arom	Arom
26	Arom	Arom	Arom	Arom	Arom
27	Arom	Arom	Arom	Arom	Arom
28	Arom	Arom	Arom	Arom	Arom
29	Arom	Arom	Arom	Arom	Arom
30	Arom	Arom	Arom	Arom	Arom
31	Arom	Arom	Arom	Arom	Arom
32	Arom	Arom	Arom	Arom	Arom
33	Arom	Arom	Arom	Arom	Arom
correct	33	32	32	33	

Table 4.8: Classification results of the four algorithms on class Naph of the first PIANO dataset.

ID	**PCA with DFA**	**SIMCA**	**Decision trees**	**MSN-PSSM**	**True class label**
1	Naph	Naph	Naph	Naph	Naph
2	Olef	Naph	Naph	Naph	Naph
3	Naph	Naph	Naph	Naph	Naph
4	Naph	Naph	Naph	Naph	Naph
5	Naph	Naph	Naph	Naph	Naph
6	Naph	Naph	Naph	Naph	Naph
7	Isopara	Naph	Naph	Naph	Naph
8	Naph	Naph	Naph	Naph	Naph
9	Olef	Naph	Naph	Naph	Naph
10	Naph	Naph	Naph	Naph	Naph
11	Olef	Naph	Olef	Naph	Naph
12	Naph	Naph	Naph	Naph	Naph
13	Olef	Naph	Naph	Naph	Naph
14	Naph	Isopara	Naph	Naph	Naph
15	Naph	Naph	Olef	Naph	Naph
16	Naph	Isopara	Naph	Naph	Naph
17	Olef	Isopara	Isopara	Naph	Naph
18	Olef	Isopara	Olef	Naph	Naph
19	Isopara	Isopara	Isopara	Isopara	Naph
20	Olef	Naph	Naph	Olef	Naph
correct	11	15	15	18	

Table 4.9: Classification results of the four algorithms on class Olef of the first PIANO dataset.

ID	PCA with DFA	SIMCA	Decision trees	MSN-PSSM	True class label
1	Olef	Olef	Naph	Olef	Olef
2	Olef	Olef	Naph	Olef	Olef
3	Olef	Olef	Olef	Olef	Olef
4	Olef	Naph	Naph	Naph	Olef
5	Olef	Isopara	Olef	Olef	Olef
6	Olef	Isopara	Naph	Naph	Olef
7	Olef	Isopara	Olef	Naph	Olef
8	Naph	Isopara	Naph	Naph	Olef
9	Naph	Isopara	Naph	Naph	Olef
10	Naph	Isopara	Olef	Naph	Olef
11	Olef	Isopara	Isopara	Naph	Olef
correct	8	3	4	4	

racy of 89.29%. SIMCA and PCA with DFA have similar overall classification accuracies of 83.93% and 82.74%. Decision tree learning has 73.81% overall classification accuracy. The performance of these four classification algorithms would be achieved by random guessing (without a priori knowledge of class distribution) with less than 1.0×10^{-16} probability (greater than 99.9999% significance level) based on the binomial test of significance. The spectra's predicted class labels of MSN-PSSM have almost perfect agreement with the spectra's true class labels based on Table 4.1. The spectra's predicted class labels of the other three supervised classification algorithms hold only substantial agreement with the spectra's true class labels based on Table 4.1. The MSN-PSSM algorithm significantly outperforms the SIMCA algorithm at 89.38% significance level based on the paired t-test. MSN-PSSM significantly outperforms PCA with DFA at 97.26% significance level, and significantly outperforms decision tree learning at 99.99% significance level.

Table 4.11 shows the confusion matrix, the precision of each class, and the recall of each

class for each supervised classification algorithm on the second PIANO dataset. Table 4.11 shows MSN-PSSM can successfully discriminate between PIANO categories, and MSN-PSSM outperforms PCA with DFA, SIMCA, and decision tree learning.

For class Para, the MSN-PSSM algorithm successfully classifies most spectra correctly, with five of forty spectra misclassified. PCA with DFA and decision tree learning have the same performance of misclassifying ten spectra. SIMCA performs relatively poorly with seventeen spectra misclassified. Nearly half misclassified spectra are classified as class Isopara, which is consistent with the close structural similarity between class Para and class Isopara.

For class Isopara, SIMCA successfully classifies most spectra correctly, with two of thirty-five spectra misclassified. PCA with DFA and MSN-PSSM have similar performance, with five and seven misclassified spectra respectively. Decision tree learning has the worst performance with eleven misclassified spectra. Most misclassified spectra are classified as class Para, which is consistent with the close structural similarity between class Isopara and class Para.

For class Arom, MSN-PSSM and decision tree learning both successfully classify all thirty-nine data correctly. They are able to perfectly discriminate class Arom from other

Table 4.10: Performance of classifiers on the second PIANO dataset. Boldface indicates the best performance.

Classifier	**Accuracy (%)**	**Kappa**
PCA with DFA	82.74	0.78
SIMCA	83.93	0.80
Decision trees	73.81	0.67
MSN-PSSM	**89.29**	**0.87**

Table 4.11: Confusion matrix, precision, and recall of each classification algorithm on the second PIANO dataset.

Classifier	**Confusion matrix**							
PCA with DFA		Para	Isopara	Arom	Naph	Olef	error	recall
	Para	30	10	0	0	0	10	0.75
	Isopara	5	30	0	0	0	5	0.86
	Arom	0	1	38	0	0	1	0.97
	Naph	0	0	0	18	8	8	0.69
	Olef	0	1	0	4	23	5	0.82
	error	5	12	0	4	8	29	
	precision	0.86	0.71	1.00	0.82	0.74		
SIMCA		Para	Isopara	Arom	Naph	Olef	error	recall
	Para	23	16	0	1	0	17	0.58
	Isopara	2	33	0	0	0	2	0.94
	Arom	0	0	38	1	0	1	0.97
	Naph	0	3	0	23	0	3	0.88
	Olef	0	2	0	2	24	4	0.86
	error	2	21	0	4	0	27	
	precision	0.92	0.61	1.00	0.85	1.00		
Decision trees		Para	Isopara	Arom	Naph	Olef	error	recall
	Para	30	8	1	1	0	10	0.75
	Isopara	9	24	0	1	1	11	0.69
	Arom	0	0	39	0	0	0	1.00
	Naph	0	3	0	10	13	16	0.38
	Olef	0	1	0	6	21	7	0.75
	error	9	12	1	8	14	44	
	precision	0.77	0.67	0.98	0.56	0.60		
MSN-PSSM		Para	Isopara	Arom	Naph	Olef	error	recall
	Para	35	5	0	0	0	5	0.88
	Isopara	2	28	0	5	0	7	0.80
	Arom	0	0	39	0	0	0	1.00
	Naph	0	0	0	25	1	1	0.96
	Olef	0	0	0	5	23	5	0.82
	error	2	5	0	10	1	18	
	precision	0.95	0.85	1.00	0.71	0.96		

classes. SIMCA and PCA with DFA have only one misclassified spectrum. The fact that class Arom is easy to discriminate from other classes is consistent with the unique benzene ring structure of class Arom.

For class Naph, MSN-PSSM and SIMCA have similar performance of successfully classifying most of the twenty-six spectra correctly. MSN-PSSM has only one misclassified spectrum. SIMCA has three misclassified spectra. PCA with DFA has worse performance, with eight misclassified spectra. Decision tree learning has the worst performance, with sixteen misclassified spectra.

For class Olef, the SIMCA algorithm has the best performance with four misclassified spectra of twenty-eight spectra. MSN-PSSM and PCA with DFA have the same performance of correctly classifying twenty-three spectra and misclassifying five spectra. Decision tree learning has the worst performance compared with SIMCA, MSN-PSSM, and PCA with DFA. Decision tree learning misclassifies seven spectra.

The first PIANO dataset was acquired with a quadrupole GC$\times$GC-MS instrument, and carries unavoidable instrument noise and chemical noise. The second PIANO dataset was acquired from the NIST/EPA/NIH Mass Spectral Library 2005 (NIST05), and carries much less instrument noise and chemical noise. The success of the MSN-PSSM algorithm on both of these two datasets and the fact that MSN-PSSM consistently outperforms other algorithms on both of these two datasets show the robustness of the algorithm.

4.5.3 UTI Dataset

Table 4.12 shows the overall classification accuracy and Fleiss kappa statistic of each supervised classification algorithm on the UTI dataset. The MSN-PSSM algorithm outperforms the other three algorithms with the highest overall classification accuracy of 70.14%. SIMCA and decision tree learning have similar overall classification accuracies of 45.14% and 47.92%. PCA with DFA has 56.94% overall classification accuracy which is better than SIMCA and decision tree learning. The performance of these four classification algorithms would be achieved by random guessing (without a priori knowledge of class distribution) with less than 1.0×10^{-16} probability (greater than 99.9999% significance level) based on the binomial test of significance. The spectra's predicted class labels of MSN-PSSM have substantial agreement with the spectra's true class labels based on Table 4.1. The spectra's predicted class labels of the other three supervised classification algorithms hold only moderate agreement with the spectra's true class labels based on Table 4.1. The MSN-PSSM algorithm significantly outperforms the PCA with DFA algorithm at 98.55% significance level based on the paired t-test. MSN-PSSM significantly outperforms SIMCA at 99.9999% significance level, and outperforms decision tree learning at 99.9989% significance level.

Table 4.12: Performance of classifiers on the UTI dataset. Boldface indicates the best performance.

Classifier	**Accuracy (%)**	**Kappa**
PCA with DFA	56.94	0.54
SIMCA	45.14	0.41
Decision trees	47.92	0.44
MSN-PSSM	**70.14**	**0.68**

Table 4.13 shows the confusion matrix, the precision of each class, and the recall of each class for the MSN-PSSM algorithm. The MSN-PSSM algorithm successfully classifies all data of strain Esp3 correctly. It classifies most of the data (6 to 8 of 9) correctly for strains Cfr1, Cfr2, Eco2, Eco3, Eco5, Esp1, Esp2, Esp4, Kpn3, and Pmi. It classifies about half (4 or 5 of 9) of the data correctly for strains Eco1, Eco4, Kox, Kpn1, and Kpn2. Table 4.13 shows MSN-PSSM can successfully discriminate between bacterial strains.

Table 4.14 shows the confusion matrix, the precision of each class, and the recall of each class for PCA with DFA on the UTI dataset. PCA with DFA successfully classifies all data of strain Kpn2 correctly. It classifies most of the data (6 to 8 of 9) correctly for strains Cfr2, Eco5, Esp1, Esp2, Esp4, and Kpn3. It classifies about half (4 or 5 of 9) of the data correctly for strains Cfr1, Eco2, Eco4, Esp3, Kox, and Kpn1. It incorrectly classifies most of the data (6 to 8 of 9) for strains Eco1, Eco3, and Pmi.

Table 4.15 shows the confusion matrix, the precision of each class, and the recall of each class for SIMCA on the UTI dataset. SIMCA classifies most of the data (6 to 8 of 9) correctly for strains Cfr2, Esp3, Kpn1, and Kpn3. It classifies about half (4 or 5 of 9) of the data correctly for strains Cfr1, Eco3, Eco4, Esp2, Kox, Kpn2, and Pmi. It incorrectly classifies most of the data (6 to 8 of 9) for strains Eco1, Eco2, and Esp1. It misclassifies all data of strain Eco5 and Esp4.

Table 4.16 shows the confusion matrix, the precision of each class, and the recall of each class for decision tree learning on the UTI dataset. Decision tree learning successfully classifies all data of strain Esp1 correctly. It classifies most of the data (6 to 8 of 9) correctly for strains Cfr2, Esp3, and Kox. It classifies about half (4 or 5 of 9) of the data correctly for strains Eco2, Eco3, Eco5, Esp2, Kpn1, Kpn2, and Pmi. It incorrectly classifies most of the data (6 to 8 of 9) for strains Cfr1, Eco1, Eco4, Esp4, and Kpn3.

Table 4.13: Confusion matrix, precision, and recall of the MSN-PSSM algorithm on the UTI dataset.

Classifier	**Confusion matrix**																	
		Cfr1	Cfr2	Eco1	Eco2	Eco3	Eco4	Eco5	Esp1	Esp2	Esp3	Esp4	Kox	Kpn1	Kpn2	Kpn3	Pmi	error	recall
	Cfr1	8	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.89
	Cfr2	0	6	0	0	0	1	0	0	0	0	0	0	0	0	0	2	3	0.67
	Eco1	3	0	4	0	0	0	2	0	0	0	0	0	0	0	0	0	5	0.44
	Eco2	0	0	0	7	0	1	0	0	0	0	0	0	1	0	0	0	2	0.78
	Eco3	0	0	0	0	6	1	0	0	0	0	0	0	0	0	0	2	3	0.67
	Eco4	0	0	0	3	1	4	0	0	0	0	0	0	1	0	0	0	5	0.44
	Eco5	0	0	1	0	1	0	7	0	0	0	0	0	0	0	0	0	2	0.78
MSN-PSSM	Esp1	1	0	0	0	0	0	0	7	1	0	0	0	0	0	0	0	2	0.78
	Esp2	0	0	0	0	0	0	0	0	8	0	1	0	0	0	0	0	1	0.89
	Esp3	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	1.00
	Esp4	0	0	0	0	0	0	0	0	2	0	7	0	0	0	0	0	2	0.78
	Kox	3	0	0	0	1	0	0	0	0	0	0	4	0	0	0	1	5	0.44
	Kpn1	1	0	0	1	0	0	0	0	0	0	0	0	5	1	0	1	4	0.56
	Kpn2	0	0	0	0	0	0	0	0	0	0	0	0	4	5	0	0	4	0.56
	Kpn3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	2	2	0.78
	Pmi	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	2	0.78
	error	10	0	2	4	3	3	2	0	3	0	1	0	6	1	0	8	43	
	precision	0.44	1.00	0.67	0.64	0.67	0.57	0.78	1.00	0.73	1.00	0.88	1.00	0.45	0.83	1.00	0.47		

Table 4.14: Confusion matrix, precision, and recall of PCA with DFA on the UTI dataset.

Classifier	**Confusion matrix**																	
		Cfr1	Cfr2	Eco1	Eco2	Eco3	Eco4	Eco5	Esp1	Esp2	Esp3	Esp4	Kox	Kpn1	Kpn2	Kpn3	Pmi	error	recall
	Cfr1	4	0	0	0	0	3	0	1	1	0	0	0	0	0	0	0	5	0.44
	Cfr2	0	6	0	0	1	2	0	0	0	0	0	0	0	0	0	0	3	0.67
	Eco1	0	0	2	0	1	0	6	0	0	0	0	0	0	0	0	0	7	0.22
	Eco2	0	1	0	5	1	1	1	0	0	0	0	0	0	0	0	0	4	0.56
	Eco3	1	0	0	2	2	4	0	0	0	0	0	0	0	0	0	0	7	0.22
	Eco4	1	2	0	1	0	5	0	0	0	0	0	0	0	0	0	0	4	0.56
PCA	Eco5	0	0	0	0	1	0	8	0	0	0	0	0	0	0	0	0	1	0.89
with	Esp1	0	0	0	0	0	0	0	7	2	0	0	0	0	0	0	0	2	0.78
DFA	Esp2	0	0	0	0	0	0	0	0	6	0	3	0	0	0	0	0	3	0.67
	Esp3	0	0	0	0	0	0	0	0	3	4	2	0	0	0	0	0	5	0.44
	Esp4	0	0	0	0	0	0	0	0	0	2	7	0	0	0	0	0	2	0.78
	Kox	0	0	0	1	0	1	0	1	1	0	0	5	0	0	0	0	4	0.56
	Kpn1	0	0	0	0	0	0	0	0	0	0	0	0	4	4	1	0	5	0.44
	Kpn2	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	1.00
	Kpn3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	7	1	2	0.78
	Pmi	0	0	0	0	0	3	0	0	0	1	1	0	0	3	0	1	8	0.11
	error	2	3	0	4	4	14	7	2	7	3	6	0	0	8	1	1	62	
	precision	0.67	0.67	1.00	0.56	0.33	0.26	0.53	0.78	0.46	0.57	0.54	1.00	1.00	0.53	0.88	0.50		

Table 4.15: Confusion matrix, precision, and recall of SIMCA on the UTI dataset.

Classifier	**Confusion matrix**																	
		Cfr1	Cfr2	Eco1	Eco2	Eco3	Eco4	Eco5	Esp1	Esp2	Esp3	Esp4	Kox	Kpn1	Kpn2	Kpn3	Pmi	error	recall
	Cfr1	5	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0	4	0.56
	Cfr2	0	7	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2	0.78
	Eco1	2	0	2	0	0	4	0	1	0	0	0	0	0	0	0	0	7	0.22
	Eco2	0	0	0	2	0	5	0	0	0	0	0	2	0	0	0	0	7	0.22
	Eco3	0	0	0	0	5	0	0	0	0	0	0	2	0	0	1	1	4	0.56
	Eco4	0	0	0	2	0	5	0	0	0	0	0	2	0	0	0	0	4	0.56
	Eco5	0	0	0	0	0	7	0	0	0	0	0	0	0	1	1	0	9	0.00
SIMCA	Esp1	1	0	1	0	0	0	0	3	0	0	0	4	0	0	0	0	6	0.33
	Esp2	0	0	1	0	0	0	0	0	4	0	3	1	0	0	0	0	5	0.44
	Esp3	0	0	0	0	0	0	0	0	3	6	0	0	0	0	0	0	3	0.67
	Esp4	0	0	0	1	0	0	0	0	8	0	0	0	0	0	0	0	9	0.00
	Kox	0	0	0	0	0	2	0	0	0	0	0	5	0	0	2	0	4	0.56
	Kpn1	0	0	1	2	0	0	0	0	0	0	0	0	6	0	0	0	3	0.67
	Kpn2	0	0	0	1	0	0	1	0	0	0	0	0	2	5	0	0	4	0.56
	Kpn3	0	0	3	0	0	0	0	0	0	0	0	0	0	0	6	0	3	0.67
	Pmi	1	0	0	0	0	0	1	0	0	0	0	1	0	0	2	4	5	0.44
	error	4	0	6	6	0	18	2	2	11	1	3	14	3	2	6	1	79	
	precision	0.56	1.00	0.25	0.25	1.00	0.22	0.00	0.60	0.27	0.86	0.00	0.26	0.67	0.71	0.50	0.80		

Table 4.16: Confusion matrix, precision, and recall of decision tree learning on the UTI dataset.

Classifier	**Confusion matrix**																	
		Cfr1	Cfr2	Eco1	Eco2	Eco3	Eco4	Eco5	Esp1	Esp2	Esp3	Esp4	Kox	Kpn1	Kpn2	Kpn3	Pmi	error	recall
Decision trees	Cfr1	3	0	1	0	0	3	1	0	0	0	0	0	0	0	0	1	6	0.33
	Cfr2	0	8	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0.89
	Eco1	0	1	1	1	1	0	4	0	0	0	0	0	0	0	1	0	8	0.11
	Eco2	0	0	0	4	2	0	0	0	1	1	0	0	0	0	1	0	5	0.44
	Eco3	2	0	0	0	4	1	0	0	0	0	0	1	0	0	1	0	5	0.44
	Eco4	0	1	0	0	1	2	0	0	0	0	0	2	0	2	1	0	7	0.22
	Eco5	2	0	1	0	1	0	5	0	0	0	0	0	0	0	0	0	4	0.56
	Esp1	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	1.00
	Esp2	0	0	1	0	0	0	0	0	4	0	3	1	0	0	0	0	5	0.44
	Esp3	0	0	0	0	0	0	0	0	2	6	0	0	0	0	1	0	3	0.67
	Esp4	0	0	0	0	0	0	0	1	1	4	3	0	0	0	0	0	6	0.33
	Kox	0	0	0	0	0	0	1	1	0	0	0	6	0	0	1	0	3	0.67
	Kpn1	0	0	0	0	0	1	0	0	0	0	0	0	4	4	0	0	5	0.44
	Kpn2	0	0	0	0	0	1	0	0	0	0	0	0	4	4	0	0	5	0.44
	Kpn3	0	2	0	0	2	1	0	0	0	0	2	1	0	0	1	0	8	0.11
	Pmi	0	0	0	0	0	1	0	0	1	1	0	0	0	0	1	5	4	0.56
	error	4	4	3	1	7	8	6	2	5	6	5	6	4	6	7	1	75	
	precision	0.43	0.67	0.25	0.80	0.36	0.20	0.45	0.82	0.44	0.50	0.38	0.50	0.50	0.40	0.13	0.83		

4.6 Summary

Experimental results demonstrate the effectiveness of the new MSN-PSSM algorithm. MSN-PSSM successfully captures the difference between predefined chemical or biological classes and successfully classifies most of the test data correctly. It significantly outperforms popular techniques PCA with DFA, SIMCA, and decision tree learning.

Experimental results also demonstrate the robustness of the new MSN-PSSM algorithm. Datasets from mass spectrometers always carry noise, such as chemical noise and instrument noise. The new MSN-PSSM algorithm models the intra-class variability and uses a smoothing model in the similarity measure to enhance the robustness of handling noise. MSN-PSSM successfully applies to GC$\times$GC-MS and ToF-SIMS.

Chapter 5

Non-Targeted Cross-Sample Classification

This chapter presents a new non-targeted cross-sample classification method to analyze comprehensive two-dimensional chromatograms [28, 29].

Given:

1. n labeled comprehensive two-dimensional chromatograms $\{x_i \mid i = 1, 2, \ldots, n\}$;
2. p predefined classes $\{c_j \mid j = 1, 2, \ldots, p\}$, where c_j represents the class name of class j;
3. class labels of the n labeled chromatograms $\{y_i \mid i = 1, 2, \ldots, n\}$, where $y_i \in \{c_j \mid j = 1, 2, \ldots, p\}$;
4. a query chromatogram x_q, and $x_q \notin \{x_i \mid i = 1, 2, \ldots, n\}$;

the steps of the non-targeted cross-sample classification are:

1. Process the n labeled comprehensive two-dimensional chromatograms $\{x_i \mid i = 1, 2, \ldots, n\}$ to detect all peaks and represent those peaks in n templates $\{t_i \mid i = 1, 2, \ldots, n\}$.

2. Create a registration template with registration peaks that are correspond (*i.e.*, are matched) across chromatograms.

3. Create a cumulative chromatogram by aligning (registering) the individual chromatograms $\{x_i \mid i = 1, 2, \ldots, n\}$ using the registration template and summing the aligned chromatograms.

4. Generate a feature template by adding retention-time regions of all peaks detected in the cumulative chromatogram to the registration template.

5. Create a cross-sample feature vector that characterizes the detector response within each retention-time region for each chromatogram.

6. Build classification models for predefined classes $\{c_j \mid j = 1, 2, \ldots, p\}$ based on the cross-sample feature vectors, predict the class label of the query chromatogram x_q, and identify potential biomarkers of predefined classes for closer examination based on the discriminating features of the classification models.

The non-targeted cross-sample classification comprehensively compares every compound, whether known or unknown, across multiple chromatograms. It provides comprehensive surveys of quantitative differences in the chemical compositions among chromatograms. This non-targeted cross-sample classification avoids the intractable problem of comprehensive cross-sample peak matching by using a few registration peaks for alignment

and peak-based retention-time regions to define comprehensive features. And the registration peaks that are correspond across chromatograms are automatically and systematically detected.

Figure 5.1 illustrates the non-targeted cross-sample classification step by step. The following sections describe the non-targeted cross-sample classification in detail.

5.1 Processing

Comprehensive two-dimensional chromatograms are presented as two-dimensional images with the x-axis (abscissa) representing the retention time in the first column and the y-axis (ordinate) representing the retention time in the second column. Each chromatogram is processed for baseline correction, peak detection, and template construction with GC Image GC×GC Software R2.1®.

5.1.1 Baseline Correction

In two-dimensional chromatograms, each individual chemical compound forms a two-dimensional cluster of pixels (a peak) with values larger than the background values (the data values in which no chemical peak is present). Under controlled conditions, the background level consists primarily of the sum of two slowly varying components: a relatively steady-state standing-current offset (characteristic of detectors) and temperature-induced column-bleed [126]. Accurate peak detection and quantification of the chemical related peaks requires subtraction of the background level from the signal.

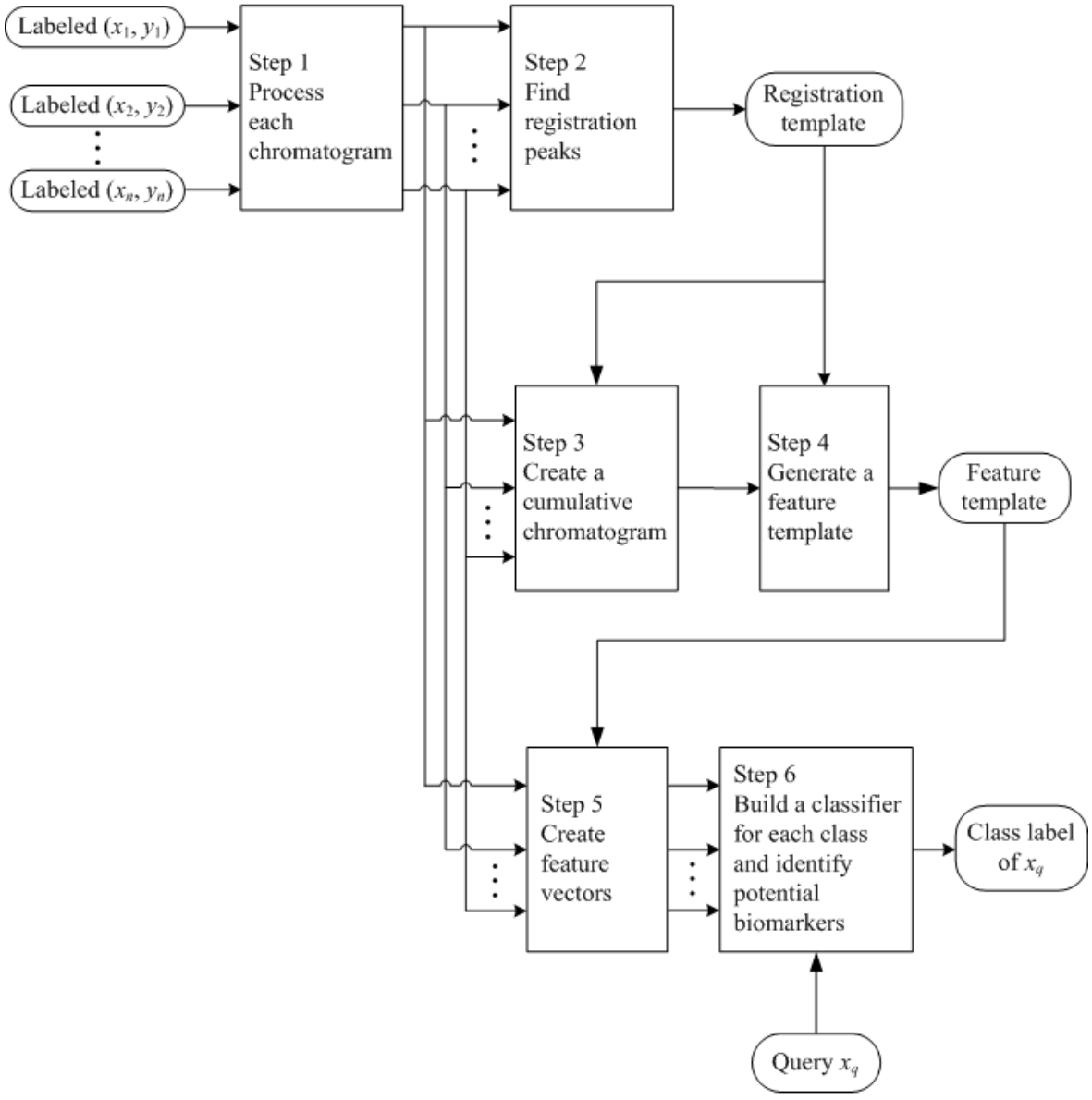

Figure 5.1: The steps of the non-targeted cross-sample classification.

Baseline correction is performed with the baseline correction algorithm developed by Reichenbach *et al*. [126]. The baseline correction algorithm estimates the baseline values across the chromatogram based on structural and statistical properties of data and then subtracts the baseline estimate from the data at each point, producing a chromatogram in which the peaks rise above a near-zero baseline. First, background regions (regions without peaks) are identified by locating data points with the smallest values in each second column chromatogram (or other interval). Then, the local means of the values from data points in the background regions are taken as first estimates of the baseline, and the variances of the values are taken as first estimates of the variance of the noise distribution (noise also is present in the background). Next, signal processing filters are used to reconstruct the baseline as a function of the local estimates. Finally, the baseline estimate is subtracted from the signal.

5.1.2 Peak Detection

The chemical peaks are detected in two dimensions using the drain algorithm [127], an inverted version of the watershed algorithm [128]. The drain algorithm is a greedy dilation algorithm that proceeds by starting peaks at tops and iteratively adding smaller pixels bordering the peaks until there are no more smaller, positive-valued pixels in the surrounds. This process can be understood conceptually by picturing the chromatograms as a relief map with larger values having higher elevation. The surface is placed under enough "water" to submerge the highest elevation; then, the water is progressively "drained". As the draining proceeds, peaks appear as "islands". As more water drains, peaks expand as lower-lying pixels around the "shore" are exposed. When the water between two peaks disappears, a border between peaks is set. This process stops when the water level reaches zero.

5.1.3 Template Construction

For each chromatogram, a template records the peak pattern of the chromatogram and captures the information for identifying the same compounds in other chromatograms. For each peak, the template records the two-dimensional retention times and a rule, expressed in the Computer Language for Identifying Chemicals (CLIC), which specifies the expected mass spectrum and the required NIST match factor [129]. The match factor required for identification is determined by analyzing the match factors with neighboring peaks (so that a peak among other peaks with similar mass spectra may require a higher match factor than a peak among other peaks with dissimilar mass spectra) [130].

5.2 Registration Template

A registration template records registration peaks that are correspond across chromatograms. The registration peaks should include peaks across the retention-time plane for chromatographic alignment.

5.2.1 Template Matching

Each template $t_i (i \in \{1, 2, \ldots, n\})$ from each chromatogram x_i is matched to all of the other chromatograms $\{x_j \mid j = 1, 2, \ldots, n \text{ and } j \neq i\}$. The matching is performed with the template matching algorithm developed by Ni and Reichenbach [131]. The template matching algorithm uses retention times and mass spectral matching rules to match a template peak from one chromatogram to at most one detected peak in each other chromatogram.

The template matching is performed for each template to each of the other chromatograms. For n chromatograms, there are $n(n-1)$ template matchings.

In a graph, the peaks can be represented as vertices and the peak matchings can be represented as directed edges. A directed edge from one vertex to another indicates a matching of a template peak from one chromatogram to a detected peak in another chromatogram. Each peak has at most $n-1$ outgoing edges, with at most one edge to each of the other chromatograms. And, each peak has at most $n-1$ incoming edges, with at most one edge from each of the other chromatograms. Figure 5.2, discussed subsequently in more detail, illustrates example matchings between a few peaks in three chromatograms x_1, x_2, and x_3.

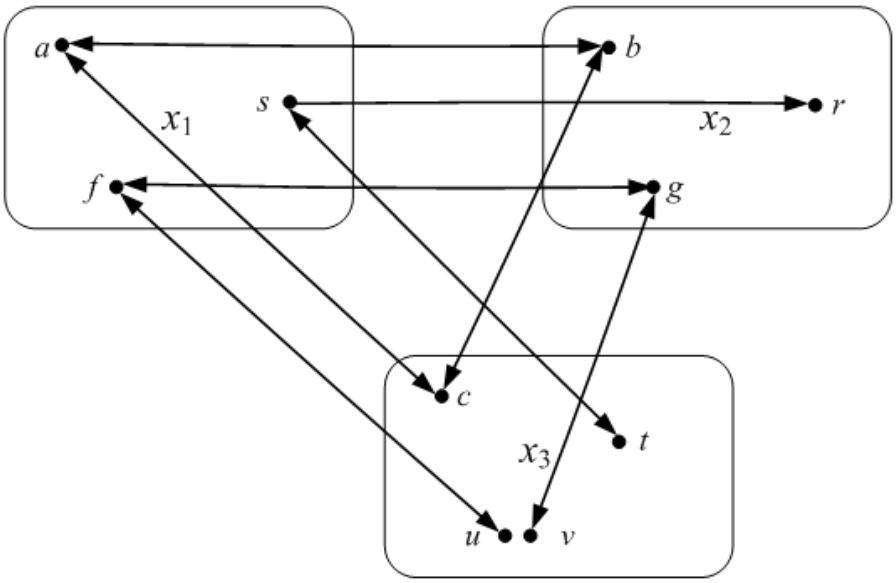

Figure 5.2: Graph visualization of example peak matching across three chromatograms x_1, x_2, and x_3.

5.2.2 Reliably Matched Peaks Detection

If peak a in the template from chromatogram x_1 matches peak b detected in chromatogram x_2 and peak b in the template from chromatogram x_2 matches peak a detected in chromatogram x_1, the peaks a and b are said to correspond. In Figure 5.2, peak a of x_1 corresponds with peak b of x_2, peak b of x_2 corresponds with peak c of x_3, and peak c of x_3 corresponds with peak a of x_1.

For a set consisting of one peak from each chromatogram, if each pair of peaks corresponds in their respective chromatograms, then the peaks are matched reliably across all chromatograms. In Figure 5.2, peaks a, b, and c are matched reliably across all three chromatograms. Peak f of x_1 corresponds with peak g of x_2 and with peak u of x_3, but peak g of x_2 and peak u of x_3 do not correspond. So these peaks are not matched reliably across all three chromatograms. Peak s of x_1 corresponds peak t of x_3, but neither has a corresponding peak in chromatogram x_2, so these peaks are not matched reliably across all three chromatograms.

In graph theory, the peaks matched reliably across all chromatograms compose a bidirectionally connected clique with n vertices, where n is the number of chromatograms. Figure 5.3 shows peaks a, b, c, d, ..., and e are matched reliably across all n chromatograms, and compose a bidirectionally connected clique of size n.

With the requirement of correspondences across all pairs of chromatograms, the number of peak matchings required for a set of reliably matched peaks is $n(n-1)$. Given $n(n \geq 2)$ chromatograms $\{x_i \mid i = 1, 2, \ldots, n\}$ and $n(n-1)$ corresponding template matchings, the steps of detecting the peaks that are matched reliably across all n chromatograms are:

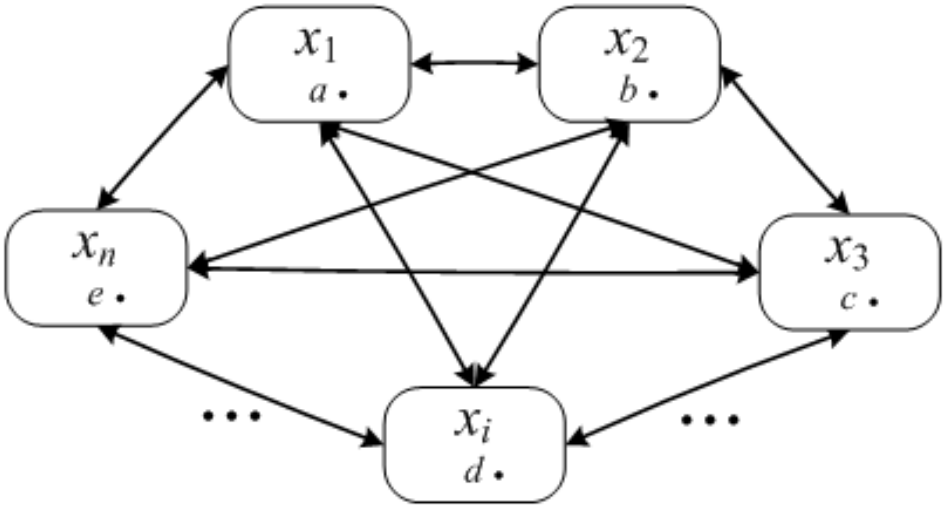

Figure 5.3: Peaks a, b, c, d, ..., and e are matched reliably across all n chromatograms, and compose a bidirectionally connected clique of size n.

1. Initialize i to be 1.

2. If peak a of chromatogram x_i matches peak b of chromatogram x_{i+1}, but peak b of chromatogram x_{i+1} does not match peak a of chromatogram x_i, report "Peaks are not matched reliably." and then exit. Otherwise, go to the next step.

3. Set r to be $i - 1$.

4. If r is greater than 0, go to the next step. Otherwise, go to step 8.

5. If peak b of chromatogram x_{i+1} matches peak d of chromatogram x_r, but peak d of chromatogram x_r does not match peak b of chromatogram x_{i+1}, report "Peaks are not matched reliably." and then exit. Otherwise, go to the next step.

6. Decrease r by 1.

7. If r is greater than 0, repeat step 5 and step 6. Otherwise, go to the next step.

8. Increase i by 1.

9. If i is less than n, repeat steps 2 to 8. Otherwise, report the peaks (one from each chromatogram) which are matched reliably across all n chromatograms.

Figure 5.4 illustrates the pseudocode implementing the above steps.

```
for i ← 1 to (n − 1) do
  if (peak a of chromatogram x_i matches peak b of chromatogram x_{i+1}) and (peak b of
  chromatogram x_{i+1} does not match peak a of chromatogram x_i) then
    report "Peaks are not matched reliably.";
    exit;
  end if
  r ← (i − 1);
  while (r > 0) do
    if (peak b of chromatogram x_{i+1} matches peak d of chromatogram x_r) and (peak d
    of chromatogram x_r does not match peak b of chromatogram x_{i+1}) then
      report "Peaks are not matched reliably.";
      exit;
    end if
    r ← (r − 1);
  end while
end for
report the peaks which are matched reliably;
exit;
```

Figure 5.4: Pseudocode of detecting the peaks which are matched reliably across all n chromatograms.

Figure 5.5 illustrates the flow chart of detecting the peaks that are matched reliably across all n chromatograms. Repeat the above steps to detect all sets of reliably matched peaks.

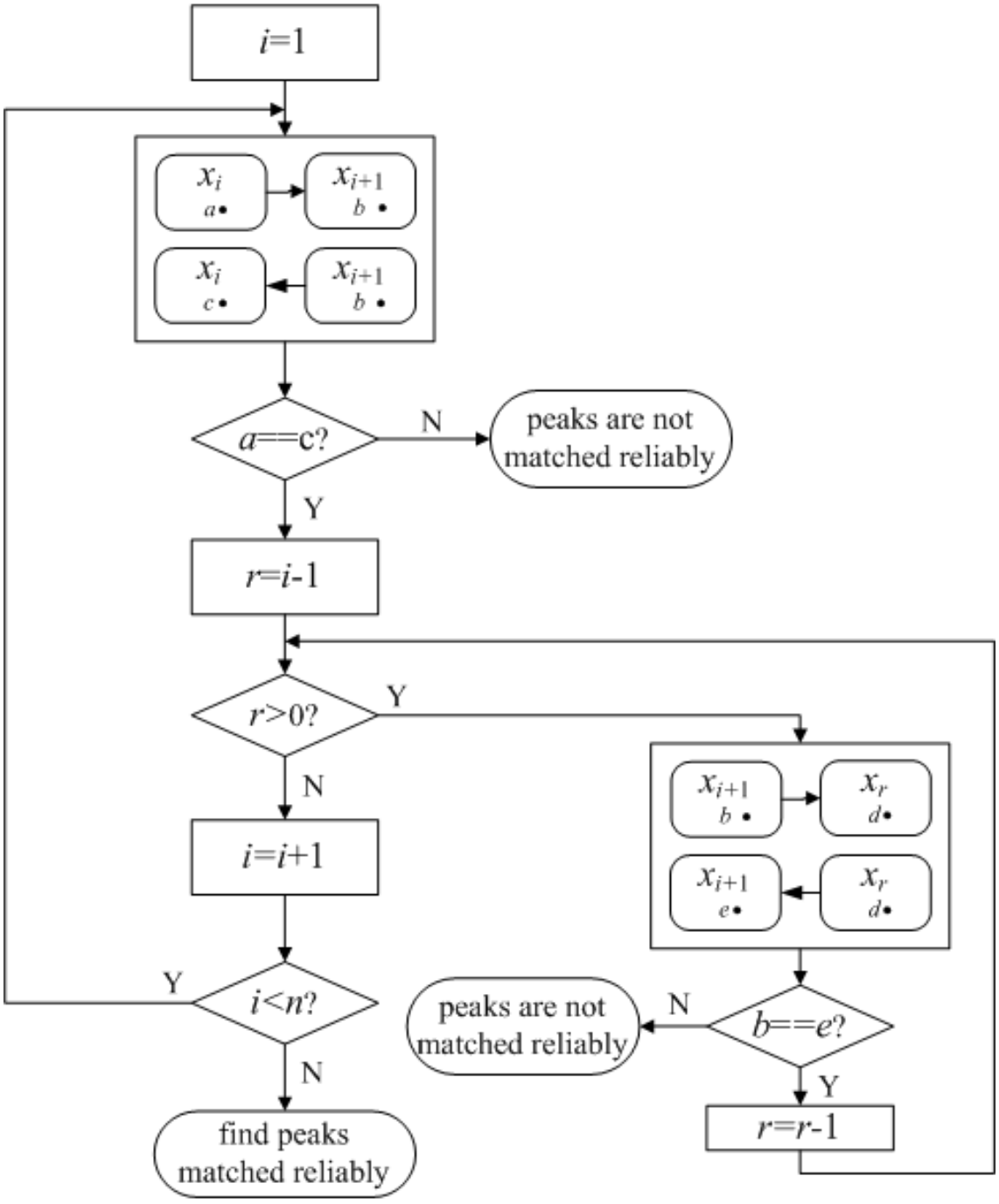

Figure 5.5: Flow chart of detecting the peaks which are matched reliably across all n chromatograms.

5.2.3 Registration Template Construction

To represent the peaks that are matched reliably across all n chromatograms, define a registration peak as the average of peaks that are matched reliably across all n chromatograms. A registration peak has an average first column retention time, an average second column retention time, an average mass spectrum, and a CLIC rule with an average NIST match factor to distinguish that peak from other peaks. For the n chromatograms, a registration template records all the registration peaks of the n chromatograms.

5.3 Cumulative Chromatogram

Align (register) the individual chromatograms using the registration template and sum the aligned chromatograms to create a cumulative chromatogram. To align each chromatogram, match the registration peaks recorded in the registration template to the detected peaks in each chromatogram. Then, align each chromatogram with the registration template using translation in the retention-time plane. This aligning is a reversal of the usual template matching operation, in which the template is transformed to align with the chromatogram. As each chromatogram is aligned with the registration template, compute the cumulative chromatogram as the pointwise sum of the individual registered chromatograms.

5.4 Feature Template

Process the cumulative chromatogram to correct the baseline and detect all peaks. For each peak detected in the cumulative chromatogram, a feature region is defined as the footprint in the retention-time plane occupied by the peak. Create a feature template by adding feature regions of all peaks detected in the cumulative chromatogram to the registration template. Figure 5.6 illustrates the feature template for the example in Figure 5.2, with one registration peak (more than one registration peak would be desired in practice) and four feature regions.

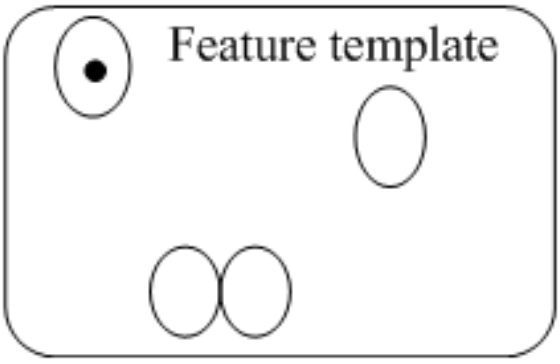

Figure 5.6: Feature template with one registration peak (filled circle) and four feature regions (open ovals).

5.5 Cross-Sample Feature Vector

For each chromatogram, a cross-sample feature vector characterizes the detector response within the regions of the feature template. Match the registration peaks of the feature template to the detected peaks in each chromatogram. The matching uses both the retention-

time plane pattern and the mass spectral matching rules of the template's registration peaks. Then, for each chromatogram, align the feature template to the chromatogram using translation and scaling. Applying the translation and scaling to both the registration peaks and feature regions in the feature template maintains the geometries of the regions relative to the registration peaks and brings them into proper alignment with the detected peaks in each chromatogram. For each chromatogram, with each transformed feature, the cross-sample feature vector value is computed as the total TIC summed over all data points in the feature region (other characteristics, such as total selected ion count values, could be used). Then, normalize the feature values for each chromatogram by dividing by the sum of all the feature values in the chromatogram so that each feature is a fractional response (other normalization methods, such as normalizing to internal standard or reference peaks, could be used).

5.6 Classification

A classifier models the predefined classes, predicts the class label of unseen chromatograms based on the feature vector of each chromatogram, and identifies potential biomarkers of predefined classes for closer examination based on the discriminating features of the classifier. The process of building the classifier attempts to determine which features are indicative of the class label and the manner in which they are indicative so that the class label of an unknown can be predicted.

Leave-one-out cross-validation is a standard accuracy estimation technique to estimate prediction accuracy. Leave-one-out cross-validation conducts the classification experiment once for each chromatogram. In each experiment, leave-one-out cross-validation partitions

the data set into a testing set with just the subject chromatogram without its class label (the class label of the test is known but not provided to the classifier) and a training set with all of the other chromatograms with their class labels. Then, leave-one-out cross-validation constructs a classifier based on the training set (according to whichever classification method used) and classifies the testing set by inputting its feature vector into the classifier (which then predicts its class label). If the predicted class label is the same as the known class label, then the classifier is credited with a correct classification. Overall classification accuracy, precision of each class, and recall of each class are used to quantitatively measure the performance of the classifier. The overall accuracy is defined as the number of chromatograms that are classified correctly divided by the number of chromatograms that are classified. The precision measures the accuracy that a specific class has been predicted. The recall measures the ability of a classification algorithm to select instances of a certain class from a data set.

This study uses the feature vectors of comprehensive two-dimensional chromatograms for classification, but other analysis (such as clustering analysis, Fisher ratio analysis, etc.) also could use these feature vectors as discussed in Chapter 7.

Chapter 6

Experimental Results for the Non-Targeted Cross-Sample Classification

Experimental results demonstrate the effectiveness of the new non-targeted cross-sample classification. The feature vectors generated by the new non-targeted cross-sample classification are useful for discriminating between samples of different classes and providing information that can be used to identify potential biomarkers for closer examination.

6.1 Dataset

The new non-targeted cross-sample classification is demonstrated with an experimental data set from breast cancer tumor samples provided by Dr. Oliver Fiehn, University of

California - Davis. The samples were obtained from breast cancer tumors from 18 individuals, six each for grades 1-3, as determined by a cancer pathologist. Extraction protocols followed Fiehn *et al*. [132]. Sample preparation was performed at Zoex Corporation (Houston TX, USA). GC×GC separations were performed by Tofwerk AG (Thun, Switzerland) on an Agilent 7890 GC and 7693 autosampler and coupled with the Zoex FasTOF time-of-flight high-resolution mass spectrometry system. Figure 6.1 illustrates the GC×GC-MS chromatograms of the grade 1 breast cancer tumors; Figure 6.2 illustrates the grade 2 breast cancer tumors; and Figure 6.3 illustrates the grade 3 breast cancer tumors.

The visualizations of Figure 6.1, Figure 6.2, and Figure 6.3 use pseudocolorization of the TIC with a cold-to-hot color scale. Chromatographic variations are visible, for example, the larger detector responses in the first sample of grade 1 (upper left in Figure 6.1) and the fourth sample of grade 3 (middle right in Figure 6.3), and the larger late-time bleed in the third sample of grade 1 (middle left in Figure 6.1) and the first sample of grade 2 (upper left in Figure 6.2). In the data for each sample, thousands of compounds are separated by GC×GC and characterized by high-resolution mass spectrometry, providing a rich source of chemical information. Comprehensive analyses of large collections of such samples may yield biochemical features that are indicative of health conditions. Such biochemical features could indicate potential bases for diagnostic tests, provide insights into disease processes, help researchers to develop more effective treatments, and give information about response to treatments.

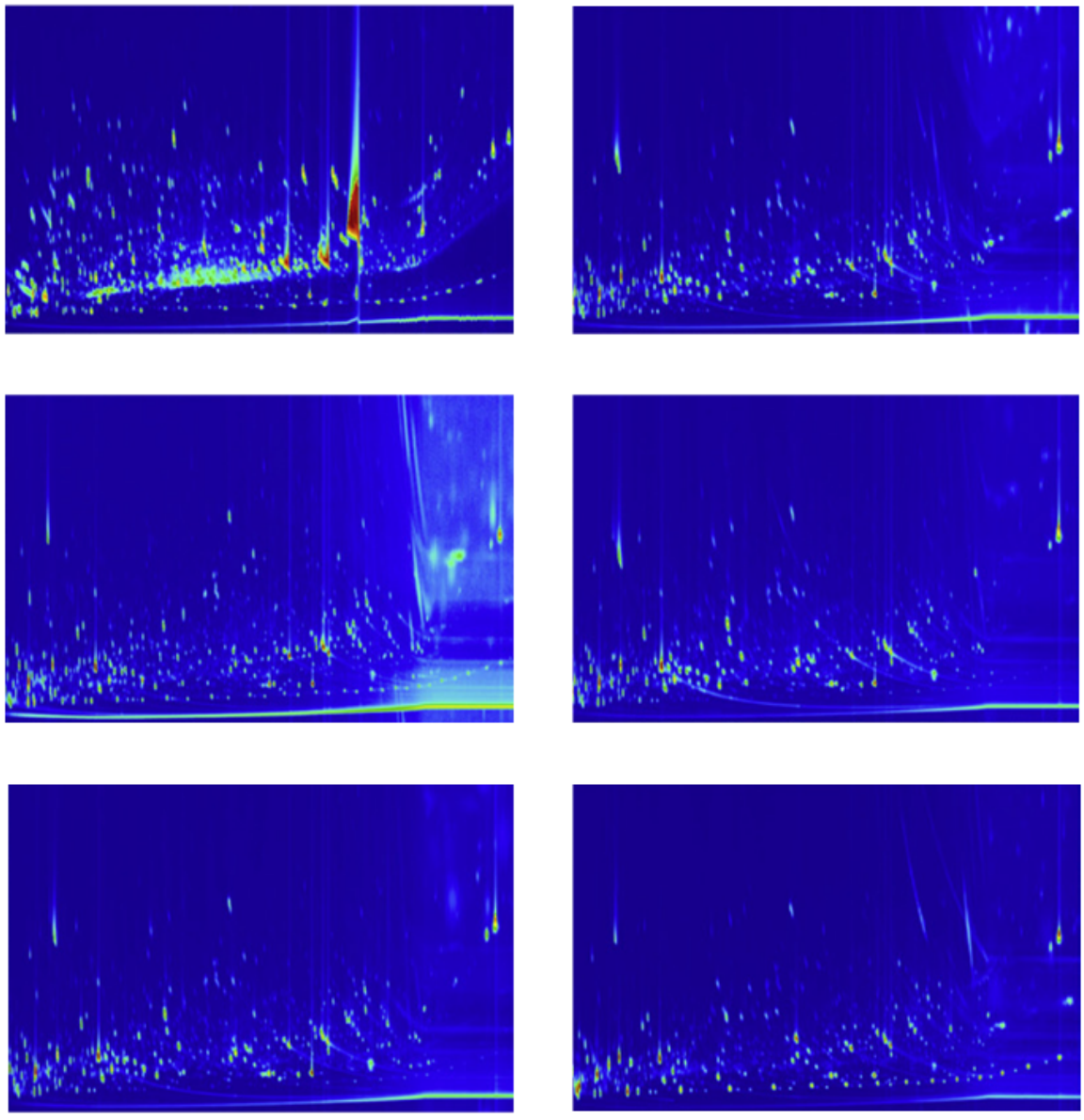

Figure 6.1: GC×GC-MS chromatograms of the grade 1 breast cancer tumors.

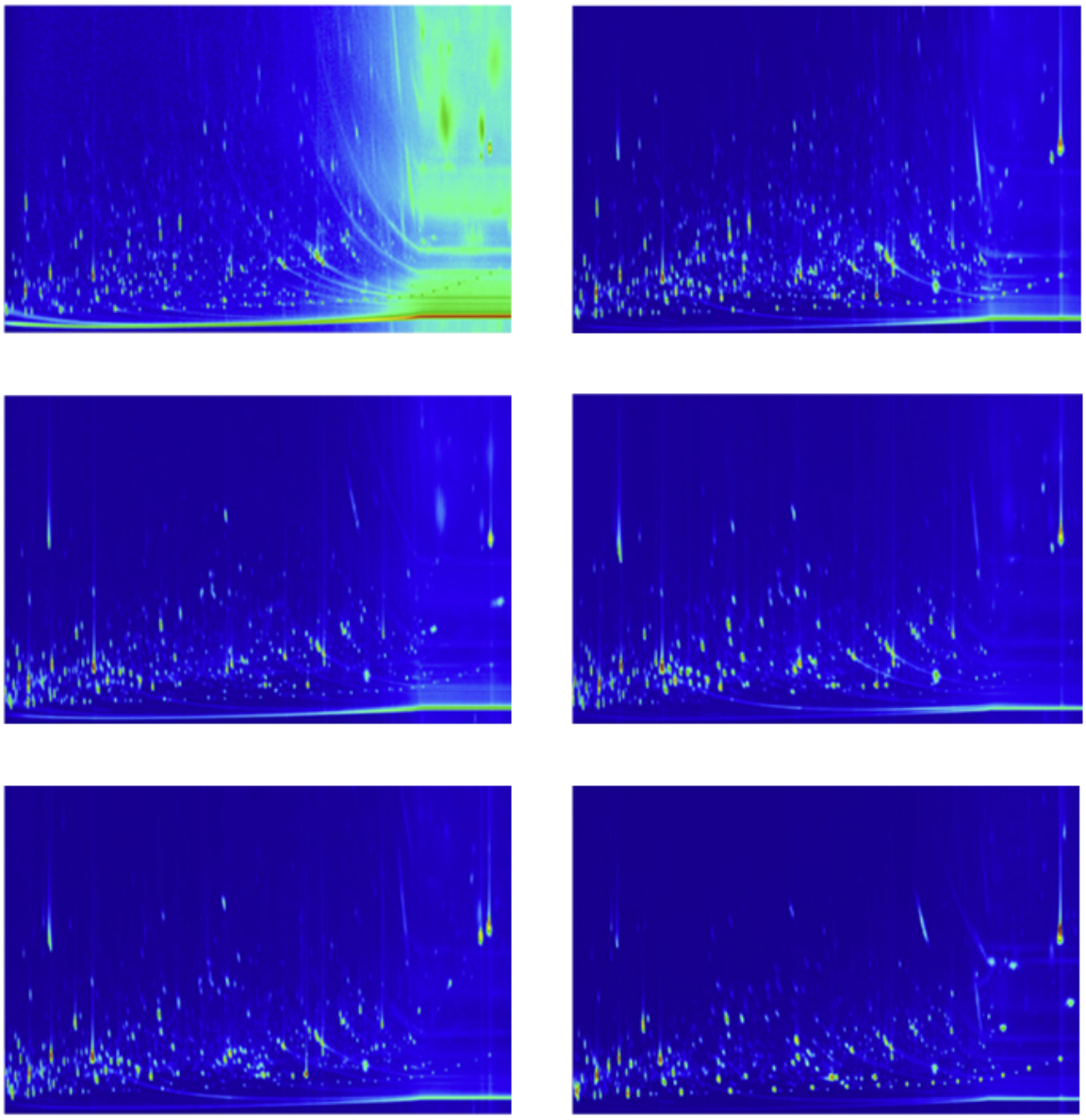

Figure 6.2: GC×GC-MS chromatograms of the grade 2 breast cancer tumors.

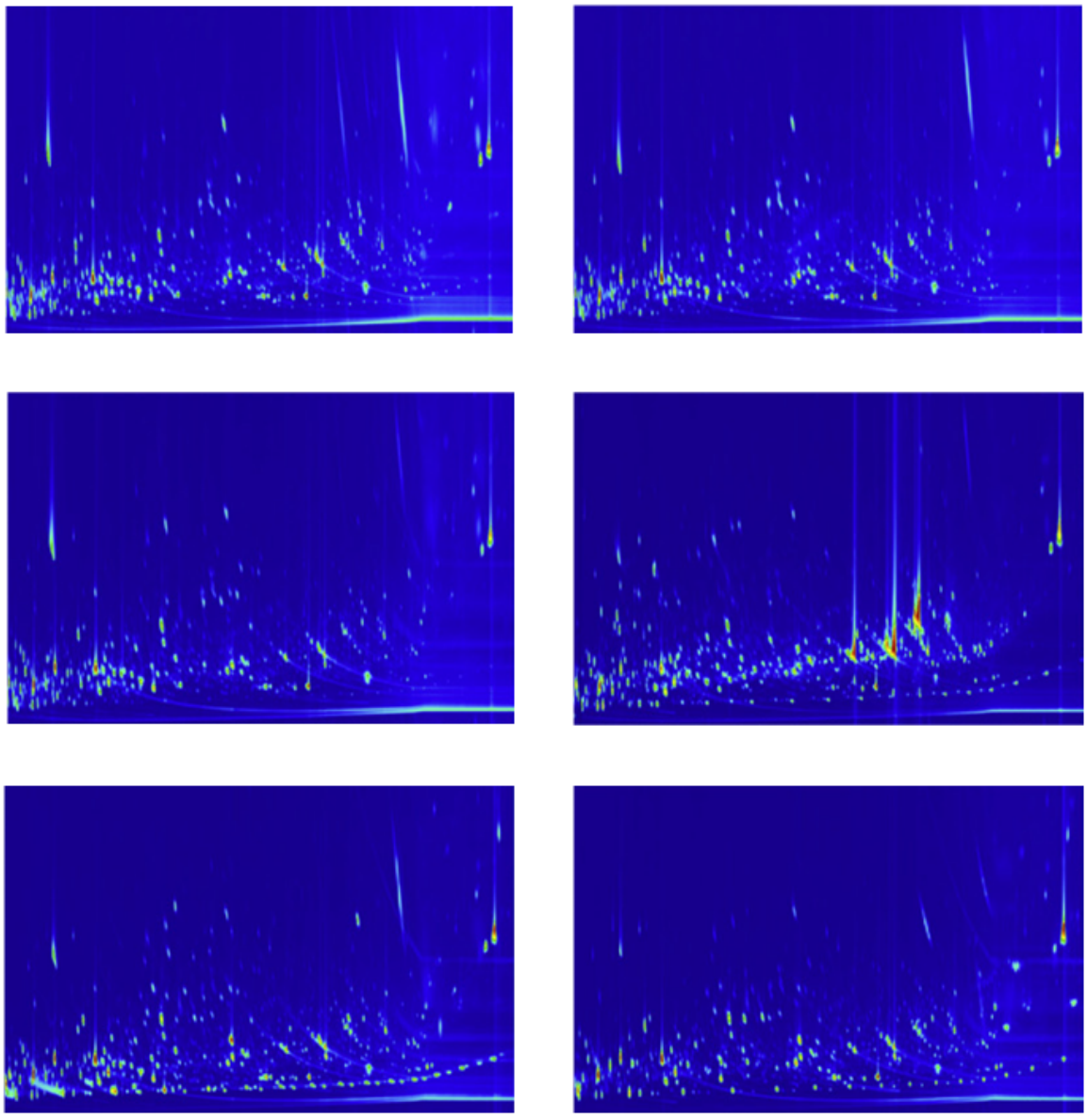

Figure 6.3: GC×GC-MS chromatograms of the grade 3 breast cancer tumors.

6.2 Experimental Results

Figure 6.4 visualizes the feature template overlaid on the cumulative chromatogram for all the samples in the breast cancer tumor data set. In the feature template, the positions of the registration peaks in the retention-time plane are annotated by black ovals. There are 13 registration peaks identified. As it can be seen, the ranges of the registration peaks nicely cover the chromatographic region in which most peaks appear. Most of the registration peaks are well separated from neighboring peaks and thus can be reliably detected and recognized across chromatograms. In the feature template, the positions of the feature regions in the retention-time plane are annotated by red outlines. There are 3408 feature regions detected.

The three grades of breast cancer indicate the degree of cellular abnormality and predict how quickly the tumor is likely to grow. The three grades are considered as three classes with class labels grade1, grade2, and grade3. Each class has six chromatograms. Without a priori knowledge of the class distribution, a classifier that guessed randomly has an expected classification accuracy of 33.33%. A classification model is built by the decision table algorithm [133] (available in the WEKA collection of machine learning algorithms [134]), which builds a table of rules based on an optimal subset of the features with wraper-based feature selection. In leave-one-out cross-validation, the decision table algorithm successfully classifies 14 chromatograms correctly and achieves a classification accuracy of 77.78%, which would be achieved by random guessing (without a priori knowledge of the class distribution) with less than 0.01% (0.0001) probability. The chromatograms' class labels generated by the decision table algorithm have substantial agreement (kappa value of 0.66) with the chromatograms' true labels according to Table 4.1. Table 6.1 shows the classification accuracy, Fleiss kappa statistics, the confusion matrix, the precision of

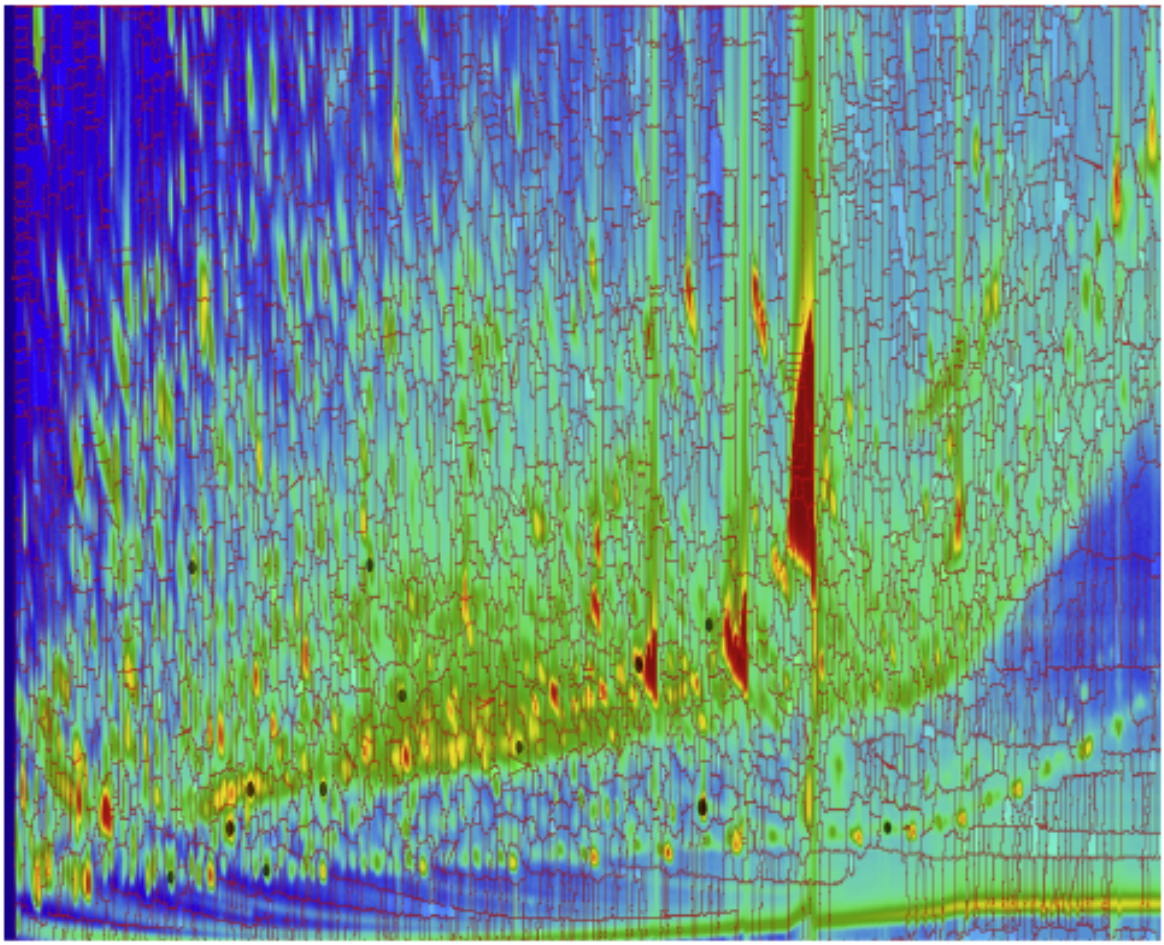

Figure 6.4: Feature template of the breast cancer data set.

each class, and the recall of each class for the decision table algorithm.

Table 6.1: Performance of decision table with leave-one-out cross-validation.

Performance	**Decision table**				
Accuracy (%)	77.78				
Kappa	0.66				
Confusion matrix		grade1	grade2	grade3	error	recall
	grade1	5	1	0	1	0.83
	grade2	1	5	0	1	0.83
	grade3	1	1	4	2	0.67
	error	2	2	0	4	
	precision	0.71	0.71	1.00		

Discriminating features identified by the classifier can be used to identify potential biomarkers for closer examination. Figure 6.5 illustrates the high-resolution mass spectrum of the region 297 which is a discriminating feature identified by decision table. Examination of the high-resolution mass spectrum peaks indicates possible elemental composition of $C_4H_{10}NOSi^+$ for the peak at mass-to-charge ratio 116 and $C_5H_{11}NOSi^+$ for the peak at mass-to-charge ratio 129, suggesting the arrangement $CHNOSi(CH_3)_3{}^+$ for the 116 fragment ion and $CHCHNOSi(CH_3)_3{}^+$ for the 129 fragment ion. Putative structure of the compound in Region 297 is illustrated in Figure 6.6 [135]. Oxime moiety is supported by exact mass measurement and elemental composition assignment [135].

6.3 Summary

Experimental results demonstrate the effectiveness of the new non-targeted cross-sample classification. The feature vectors generated by the new non-targeted cross-sample classification are useful for discriminating between breast cancer tumor samples of different

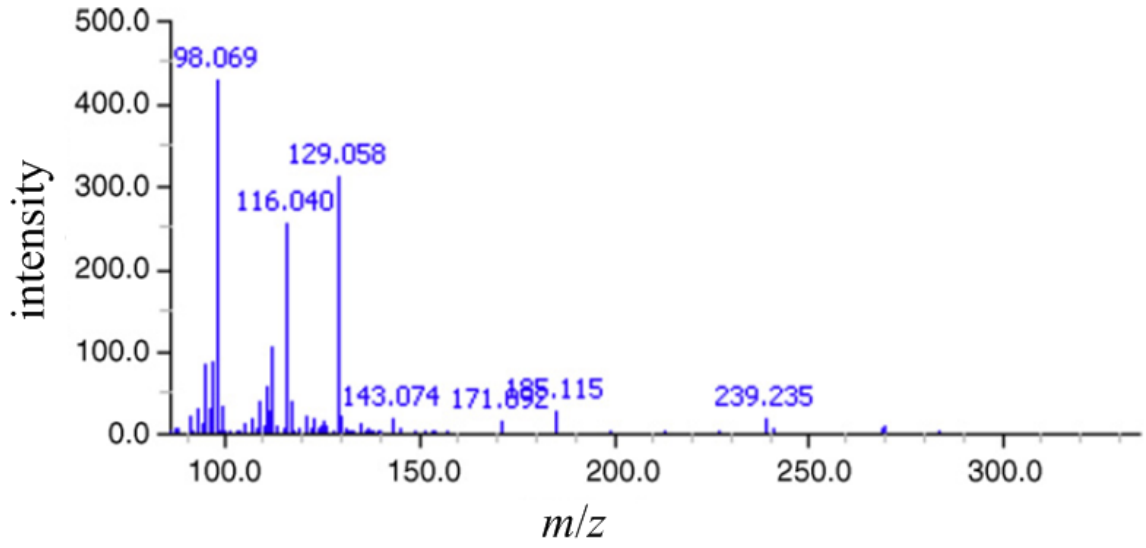

Figure 6.5: The high-resolution mass spectrum of feature 297 from one of the samples.

grades (as labeled by a cancer pathologist) and providing information that can be used to identify potential biomarkers for closer examination.

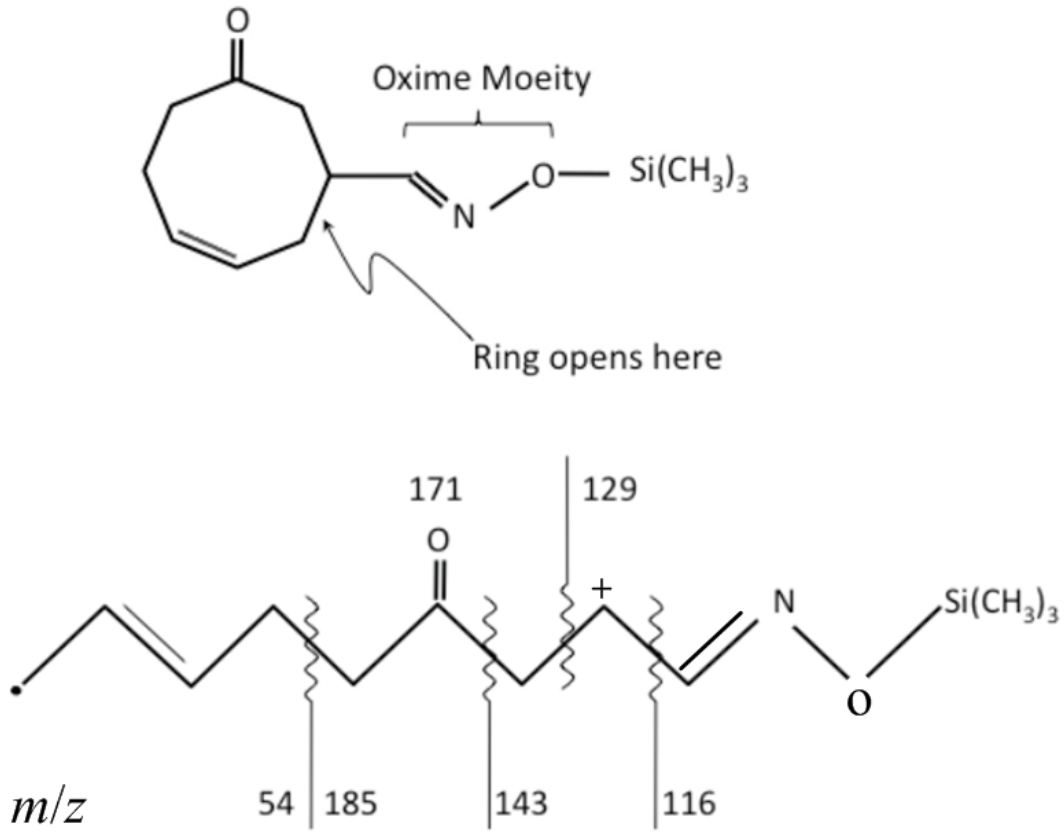

Figure 6.6: Putative structure of the compound in Region 297.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

7.1.1 Classification of Mass Spectra

Mass spectrometry is an analysis technique that measures the mass-to-charge ratio (m/z) of molecular and fragmentary ions. Mass spectra contain characteristic information regarding the composition of compounds and the properties of compounds. The mass spectra of compounds from the same chemically related group are similar. Therefore, mass spectra can be used to predict or explain compound properties, such as biological or chemical properties, based on mass spectral similarity.

Classification is one of the fundamental methodologies for analyzing mass spectral data. The primary goals of classification are to automatically group compounds based on

their mass spectra, to find correlation between the properties of compounds and their mass spectra, and to provide a positive identification of unknown compounds. Classification complements library search which searches a mass spectral library to identify unknown mass spectra. For mass spectra that cannot be found in a library, classification can involve identification of substructure types or well defined compound classes in order to establish and confirm structural conjectures or reveal relationships between mass spectra and chemical structures. Classification also can be useful in cases when only structurally related compounds need to be retrieved.

Mass spectra are high-dimensional data. Mass spectra of complex mixtures are enormously complex with large mass ranges and many structurally significant peaks combined with noise peaks (such as contaminants and small or non-diagnostic fragment ions). Within this high-dimensional complexity, there is a huge amount of information about the identity of the mixture, e.g., compound composition, molecular orientation, surface order, and chemical bonding. Establishing new mathematical or statistical methodologies for comprehensive information analysis and classification has become one of the most important tasks in mass spectral analysis.

This dissertation presents a new classification algorithm for classification of mass spectra, the most similar neighbor with a probability-based spectrum similarity measure (MSN-PSSM). The MSN-PSSM algorithm is a multi-class classification algorithm, that can deal with multiple classes directly without converting a multi-class problem into a set of two-class problems. The MSN-PSSM algorithm models the intra-class variability and uses a smoothing model in the similarity measure to enhance the robustness with respect to noise, such as chemical noise and instrument noise. The MSN-PSSM algorithm characterizes the domain information of labeled data by an array of probability distribution functions of in-

tensities as a function of m/z. Each probability in the distribution function is the fraction of spectra in the labeled data having that intensity value at the given m/z. The MSN-PSSM algorithm considers all m/z that contain discriminating information to avoid information loss.

Experimental results demonstrate the effectiveness and robustness of the new MSN-PSSM algorithm. In leave-one-out cross-validation, it outperforms popular classification techniques for classification of mass spectra, such as principal component analysis with discriminant function analysis, soft independent modeling of class analogy, and decision tree learning.

7.1.2 Cross-Sample Classification of Comprehensive Two-Dimensional Chromatograms

Two-dimensional separation patterns obtained by comprehensive chromatography, in particular comprehensive two-dimensional gas chromatography (GC×GC), analyze a complex mixture to characterize its composition. GC×GC is a powerful tool for complex biological sample characterization, differentiation, discrimination, and classification on the basis of the component distribution over the two-dimensional plane.

Comprehensive two-dimensional chromatography yields highly informative separation patterns because of its great practical peak capacity and sensitivity produced by applying two different separation principles (one for each chromatographic dimension). However, the improvement in information yields complex data (consisting of two-dimensional retention data and mass spectra) requiring comprehensive analyses to interpret the rich information and to extract useful information on sample characterization. Cross-sample analysis

of complex biological samples, such as sample classification, is even more challenging because of the difficulty of analyzing and interpreting the massive, complex data from many samples for relevant biochemical features. The large dimensionality of biological data, as well as the size of the dataset, and the possibility that significant chemical characteristics across many samples may be subtle and involve patterns of variations in multiple constituents, necessitate the investigation and development of new analysis methodologies.

This dissertation presents a new non-targeted cross-sample classification method to analyze comprehensive two-dimensional chromatograms. The non-targeted cross-sample classification systematically and automatically detects registration peaks of multiple comprehensive two-dimensional chromatograms. Then, the non-targeted cross-sample classification uses the registration peaks to align (register) the comprehensive two-dimensional chromatograms of samples to generate a cumulative chromatogram. The registration peaks and the retention-time regions of all peaks detected in the cumulative chromatogram are used to generate a feature template. The registration peaks in the feature template are matched to the detected peaks in each chromatogram. For each chromatogram, the non-targeted cross-sample classification creates a feature vector that characterizes the detector response within the regions of the feature template. Then, the non-targeted cross-sample classification uses the feature vectors for the set of comprehensive two-dimensional chromatograms to perform classification and potential biomarker identification.

The new non-targeted cross-sample classification is successfully applied to a set of comprehensive two-dimensional chromatograms of breast cancer tumor samples, each from different individuals, for cancer grades 1 to 3 (as labeled by a cancer pathologist). Experimental results demonstrate the effectiveness of the new non-targeted cross-sample classification. The feature vectors generated by the new non-targeted cross-sample classification

are useful for discriminating between breast cancer tumor samples of different grades and providing information that can be used to identify potential biomarkers for closer examination.

7.2 Future Work

The MSN-PSSM algorithm for classification of mass spectra can be extended in two aspects.

1. To simplify the domain characterization, the MSN-PSSM algorithm assumes mass-to-charge ratio independence, that is, for a spectrum, the intensity value at one mass-to-charge ratio is independent of the intensity value at any other mass-to-charge ratio. Mass-to-charge ratio independence is not usually true. The relationships between the intensities at different mass-to-charge ratios could be considered resulting in higher-order distribution functions in domain characterization. Higher-order probability distribution functions could improve the classification performance.

2. To simplify the intra-class variability modeling, the MSN-PSSM algorithm assumes the relationship between the standard deviation of intensities and the intensity level is linear. Non-linear relationship (e.g., quadratic relationship) could improve the classification performance.

The non-targeted cross-sample classification method to analyze comprehensive two-dimensional chromatograms can be extended in five aspects.

1. In this study, the non-targeted cross-sample classification detects reliably matched

peaks across all chromatograms. Reliably matching across all chromatograms is a very strong requirement because of peak detection errors as well as the inherent ambiguity of matching. Co-eluting constituents may be detected as separate peaks in some chromatograms but as one peak in other chromatograms. The peaks of different analytes may be incorrectly matched, especially if constituents differ from sample to sample. The strong requirement of reliably matching across all chromatograms could be relaxed to reliably matching across a subset of all chromatograms.

2. In this study, the non-targeted cross-sample classification uses the feature vectors of comprehensive two-dimensional chromatograms for classification. Other data analysis (such as clustering analysis, Fisher ratio analysis, etc.) can also use these feature vectors.

3. This study applies the non-targeted cross-sample classification to two-dimensional gas chromatograms. But this non-targeted cross-sample classification should not be limited to two-dimensional gas chromatograms. It should be adaptable for use with a variety of comprehensive two-dimensional chemical separations.

4. In the non-targeted cross-sample classification, each cross-sample feature value is the total TIC summed over all data points in the feature region. Considering the mass spectrum of each feature region as the feature value may improve the accuracy of the non-targeted cross-sample classification.

5. In the non-targeted cross-sample classification, each individual chromatogram is aligned to the registration template using a global transformation. Although the non-targeted cross-sample classification uses regions as features and is robust to misalignment between samples, non-global (e.g., piecewise) transformation could decrease the chance of misalignment and improve the accuracy of cross-sample classification.

Bibliography

[1] J. V. Ryzin, *Classification and Clustering*. Academic Press, 1977.

[2] C. Dass, *Principles and Practice of Biological Mass Spectrometry*. John Wiley, 2001.

[3] J. H. Gross, *Mass Spectrometry: a Textbook*. Springer-Verlag, 2004.

[4] J. T. Watson, *Introduction to Mass Spectrometry*. Raven Press, 1985.

[5] W. Bertsch, "Two-Dimensional Gas Chromatography. Concepts, Instrumentation and Applications - Part 2: Comprehensive Two-Dimensional Gas Chromatography", *Journal of High Resolution Chromatography*, vol. 23(3), pp. 167-181, 2000.

[6] G. S. Frysinger, and R. B. Gaines, "Separation and Identification of Petroleum Biomarkers by Comprehensive Two-Dimensional Gas Chromatography", *Journal of Separation Science*, vol. 24(2), pp. 87-96, 2001.

[7] A. Smilde, R. Bro, and P. Geladi, *Multi-way Analysis with Applications in the Chemical Sciences*. John Wiley, 2004.

[8] R. G. Wilson, F. A. Stevie, and C. W. Magee, *Secondary Ion Mass Spectrometry : a Practical Handbook for Depth Profiling and Bulk Impurity Analysis*. John Wiley, 1989.

[9] A. Benninghoven, F. G. Rudenauer, and H. W. Werner, *Secondary Ion Mass Spectrometry: Basic Concepts, Instrumental Aspects, Applications, and Trends*. John Wiley, 1987.

[10] H. Oechsner, *Thin Film and Depth Profile Analysis*. Springer-Verlag, 1984.

[11] R. Gnanadesikan, *Discriminant Analysis and Clustering*. National Academy Press, 1988.

[12] D. H. Smith, "A Compound Classifier Based on Computer Analysis of Low Resolution Mass Spectral Data", *Analytical Chemistry*, vol. 44(3), pp. 536-547, 1972.

[13] S. Lloyd, and C. Grimm, "Fast Temperature-programmed Gas Chromatography-Mass Spectrometry for Food Analysis", *Journal of Chromatographic Science*, vol. 40(6), pp. 309-314, 2002.

[14] M. Scotter, L. Castle, and D. Roberts, "Estimation of Sucrose Esters (e473) in Foods Using Gas Chromatography-mass Spectrometry", *Food Additives & Contaminants*, vol. 23(6), pp. 539-546, 2006.

[15] A. Caninia, D. Alesiania, G. DArcangelob, and P. Tagliatestab, "Gas Chromatography-Mass Spectrometry Analysis of Phenolic Compounds from Carica Papaya L. Leaf", *Journal of Food Composition and Analysis*, vol. 20(7), pp. 584-590, 2007.

[16] A. Lecznar, K. Stepien, E. Chodurek, S. Kurkiewicz, L. Swiatkowska, and T. Wilczok, "Pyrolysis Gas Chromatography-Mass Spectrometry of Peroxynitrite-treated Melanins", *Journal of Analytical and Applied Pyrolysis*, vol. 70(2), pp. 457-467, 2003.

[17] P. Cap, J. Chladek, F. Pehal, M. Maly, V. Petru, P. Barnes, and P. Montuschi, "Gas Chromatography/Mass Spectrometry Analysis of Exhaled Leukotrienes in Asthmatic Patients", *Thorax*, vol. 59(6), pp. 465-470, 2004.

[18] T. P. Roddy, D. M. Cannon, Jr., S. G. Ostrowski, N. Winograd, and A. G. Ewing, "Identification of Cellular Sections with Imaging Mass Spectrometry Following Freeze Fracture", *Analytical Chemistry*, vol. 74(16), pp. 4020-4026, 2002.

[19] P. Sjovall, J. Lausmaa, H. Nygren, L. Carlsson, and P. Malmberg, "Imaging of Membrane Lipids in Single Cells by Imprint-Imaging Time-of-Flight Secondary Ion Mass Spectrometry", *Analytical Chemistry*, vol. 75(14), pp. 3429-3434, 2003.

[20] R. Jeannot, H. Sabik, E. Sauvard, T. Dagnac, and K. Dohrendorf, "Determination of Endocrine-disrupting Compounds in Environmental Samples Using Gas and Liquid Chromatography with Mass Spectrometry", *Journal of Chromatography A*, vol. 974 (1-2), pp. 143-159, 2002.

[21] S. E. Stein, "Estimating Probabilities of Correct Identification from Results of Mass Spectral Library Searches", *Journal of American Society for Mass Sectrometry*, vol. 5(4), pp. 316-323, 1994.

[22] A. Visvanathan, and S. Reichenbach, "Information Theoretic Mass Spectral Library Search for Comprehensive Two-Dimensional Gas Chromatography with Mass Spectrometry", *56th ASMS Conference on Mass Spectrometry and Allied Topics*, Denver, CO, 2008.

[23] H. Scsibrany, and K. Varmuza, "Common Substructures in Groups of Compounds Exhibiting Similar Mass Spectra", *Fresenius Journal of Analytical Chemistry*, vol. 344(4-5), pp. 220-222, 1992.

[24] D. J. Graham, M. S. Wagner, and D. G. Castner, "Information from Complexity: Challenges of ToF-SIMS Data Interpretation", *Applied Surface Science*, vol. 252(19), pp. 6860-6868, 2006.

[25] M. Eggink, W. Romero, R. J. Vreuls, H. Lingeman, W. M.A. Niessen, and H. Irth, "Development and Optimization of a System for Comprehensive Two-dimensional Liquid Chromatography with UV and Mass Spectrometric Detection for the Separation of Complex Samples by Multi-step Gradient Elution", *Journal of Chromatography A*, vol. 1188(2), pp. 216-226, 2008.

[26] C. Cordero, E. Liberto, C. Bicchi, P. Rubiolo, P. Schieberle, S. E. Reichenbach, and Q. Tao, "Profiling Food Volatiles by Comprehensive Two-dimensional Gaschromatography Coupled with Mass Spectrometry: Advanced Fingerprinting Approaches for Comprarative Analysis of the Volatile Fraction of Roasted Hazelnuts (Corylus Avellana L.) from Different Origins", *Journal of Chromatography A*, vol. 1217(37), pp. 5848-5858, 2010.

[27] X. Tian, S. Reichenbach, Q. Tao, and A. Henderson, "Classification and Cluster Analysis of Complex Time-of-Flight Secondary Ion Mass Spectrometry for Biological Samples", *International Conference on Bioinformatics, Computational Biology, Genomics and Chemoinformatics*, pp. 78-85, 2009.

[28] S. E. Reichenbach, X. Tian, Q. Tao, E. B. Ledford, Z. Wu, and O. Fiehnd, "Informatics for Cross-Sample Analysis with Comprehensive Two-Dimensional Gas Chromatography and High-Resolution Mass Spectrometry (GC×GC-HRMS)", *Talanta*, DOI:10.1016/j.talanta.2010.09.057, 2010.

[29] C. Cordero, E. Liberto, C. Bicchi, P. Rubiolo, S. E. Reichenbach, X. Tian, and Q. Tao, "Targeted and Non-Targeted Approaches for Complex Natural Sample Profiling by GC×GC-qMS", *Journal of Chromatographic Science*, vol. 48(4), pp. 251-261, 2010.

[30] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points is Space", *Philosophical Magazine Series 6*, vol. 2, pp. 559-572, 1901.

[31] H. Hotelling, "Analysis of a Complex of Statistical Variable into Principal Components", *Journal of Educational Psychology*, vol. 24, pp. 417-441, 1933.

[32] W. J. Krzanowski, *Principles of Multivariate Analysis: a User's Perspective*. Oxford University Press, Revised Ed., 1988.

[33] B. F.J. Manly, *Multivariate Statistical Methods: A Primer*. Chapman and Hall, 1986.

[34] S. Wold, and M. Sjostrom, "SIMCA: A Method for Analyzing Chemical Data in terms of Similarity and Analogy", *Chemometrics Theory and Application*, American Chemical Society Symposium Series 52, pp. 243-282, Washington, D.C., 1977.

[35] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[36] J. Allen, H. M. Davey, D. Broadhurst, J. K. Heald, J. J. Rowland, S. G. Oliver, and D. B. Kell, "High-throughput Classification of Yeast Mutants for Functional Genomics Using Metabolic Footprinting", *Nature Biotechnology*, vol. 21(6), pp. 692-696, 2003.

[37] M. S. Wagner, D. J. Graham, B. D. Ratner, and D. G. Castner, "Maximizing Information Obtained from Secondary Ion Mass Spectra of Organic Thin Films Using Multivariate Analysis", *Surface Science*, vol. 570(1-2), pp. 7897, 2004.

[38] J. S. Fletcher, A. Henderson, R. M. Jarvis, N. P. Lockyer, J. C. Vickerman, and R. Goodacre, "Rapid Discrimination of the Causal Agents of Urinary Tract Infection Using ToF-SIMS with Chemometric Cluster Analysis", *Applied Surface Science*, vol. 252(19), pp. 6869-6874, 2006.

[39] C. E. Thompson, J. Ellis, J. S. Fletcher, R. Goodacre, A. Henderson, N. P. Lockyer, and J. C. Vickerman, "ToF-SIMS Studies of Bacillus Using Multivariate Analysis with Possible Identification and Taxonomic Applications", *Applied Surface Science*, vol. 252(19), pp. 6719-6722, 2006.

[40] M. J. Baker, M. D. Brown, E. Gazi, N. W. Clarke, J. C. Vickerman, and N. P. Lockyer, "Discrimination of Prostate Cancer Cells and Non-malignant Cells Using Secondary Ion Mass Spectrometry", *Analyst*, vol. 133(2), pp. 175-179, 2008.

[41] R. Baigorri, A. M. Zamarren, M. Fuentes, G. Gonzalez-Gaitano, J. M. Garcia-Mina, G. Almendros, and F. J. Gonzalez-Vila, "Multivariate Statistical Analysis of Mass Spectra as a Tool for the Classification of the Main Humic Substances According to Their Structural and Conformational Features", *Journal of Agriculture and Food Chemistry*, vol. 56(14), pp. 5480-5487, 2008.

[42] I. T. Jolliffe, *Principal Component Analysis*. Springer, Second Ed., 2002.

[43] J. E. Jackson, *A User's Guide to Principal Components*. John Wiley, 2003.

[44] J. Shlens, "A Tutorial on Principal Component Analysis Derivation, Discussion and Singular Value Decomposition", *http://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf*, 2003.

[45] K. Varmuza, and P. Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press, 2009.

[46] P. Legendre, and L. Legendre, *Numerical Ecology*. Elsevier Science, New York, 1998.

[47] W. J. Dixon, *Biomedical Computer Programs*. University of California Press, Los Angeles, 1975.

[48] H. J. H. MacFie, C. S. Gutteridge, and J. R. Norris, "Use of Canonical Variates in Differentiation of Bacteria by Pyrolysis Gas-liquid Chromatography", *Journal of General Microbiology*, vol. 104, pp. 67-74, 1978.

[49] H. R. Lindman, *Analysis of Variance in Experimental Design*. Springer-Verlag, New York, 1992.

[50] B. G. Tabachnick, and L.S. Fidell, *Using Multivariate Statistics*. Harper Collins College Publishers, New York, 1996.

[51] W. J. Dunn, S. L. Emery, and W. G. Glen, "Preprocessing, Variable Selection, and Classification Rules in the Application of SIMCA Pattern Recognition to Mass Spectral Data", *Environmental Science and Technology*, vol. 23, pp. 1499-1505, 1989.

[52] R. D. Maesschalck, A. Candolfi, D. L. Massart, and S. Heuerding, "Decision Criteria for Soft Independent Modelling of Class Analogy Applied to Near Infrared Data", vol. 47(1), pp. 65-77, 1999.

[53] J. Wu, and X. Zhang, "A PCA Classifier and Its Application in Vehicle Detection", *Proceedings of International Joint Conference on Neural Networks*, vol. 1, pp. 600-604, 2001.

[54] K. V. Branden, and M. Hubert, "Robust Classification in High Dimensions Based on the SIMCA Method", *Chemometrics and Intelligent Laboratory Systems*, vol. 79(1-2), pp. 10-21, 2005.

[55] J. B. M. Droge, W. J. Rinsma, H. A. Van T Klooster, A. C. Tas, and J. Van Der Greef, "An Evaluation of SIMCA. Part 2 - Classification of Pyrolysis Mass Spectra of Pseudomonas and Serratia Bacteria by Pattern Recognition Using the SIMCA Classifier", *Journal of Chemometrics*, vol. 1(4), pp. 231-241, 1987.

[56] M. Sarker, W. G. Glen, L. Yin, W. J. Dunn, D. R. Scott, and S. Swanson, "Comparison of SIMCA Pattern Recognition and Library Search Identification of Hazardous Compounds from Mass Spectra", *Analytica Chimica Acta*, vol. 257(2), pp. 229-238, 1992.

[57] T. Leth, "Chemometric Analysis of Mass Spectra of Cis and Srans Fatty Acid Picolinyl Esters", *Chemistry and Materials Science*, vol. 205(2), pp. 111-115, 1997.

[58] J. I. Villegas, D. Kubicka, S. P. Reinikainen, G. Addova, R. Kubinec, T. Salmi, and D. Yu. Murzin, "Classification and Pattern Recognition of Acyclic Octenes Based on Mass Spectra", *Talanta*, vol. 72(4), pp. 1573-1580, 2007.

[59] E. S. F. Berman, L. Wu, S. L. Fortson, K. S. Kulp, D. O. Nelson, and K. JenWu, "Chemometric and Statistical Analyses of ToF-SIMS Spectra of Increasingly Complex Biological Samples", *Surface and Interface Analysis*, vol. 41(2), pp. 97-104, 2008.

[60] A. Vlahou, J. O. Schorge, B. W. Gregory, and R. L. Coleman, "Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data", *Journal of Biomedicine and Biotechnology*, vol. 2003(5), pp. 308-314, DOI:10.1155/S1110724303210032, 2003.

[61] M. K. Markey, G. D. Tourassi, and C. E. Floyd, "Decision Tree Classification of Proteins Identified by Mass Spectrometry of Blood Serum Samples from People With and Without Lung Cancer", *Proteomics*, vol. 3, DOI:10.1002/pmic.200300521, 2003.

[62] C. Engrand, J. Kissel, F. R. Krueger, P. Martin, J. Silen, L. Thirkell, R. Thomas, and K. Varmuza, "Chemometric Evaluation of Time-of-Flight Secondary Ion Mass Spectrometry Data of Minerals in the Frame of Future in situ Analyses of Cometary Material by Cosima onboard ROSETTA", *Rapid Communications in Mass Spectrometry*, vol. 20(8), pp. 1361-1368, 2006.

[63] A. Assareh, and M. H. Moradi, "Knowledge Acquisition from Mass Spectra of Blood Samples Using Fuzzy Decision Tree and Genetic Algorithm", *9th International Symposium on Signal Processing and Its Applications*, DOI:10.1109/ISSPA.2007.4555376, 2007.

[64] G. Ge, and G. W. Wong, "Classification of Premalignant Pancreatic Cancer Mass-Spectrometry Data Using Decision Tree Ensembles", *BMC Bioinformatics*, vol. 9(275), DOI:10.1186/1471-2105-9-275, 2008.

[65] D. L. Swaney, G. C. McAlister, and J. J. Coon, "Decision Tree-Driven Tandem Mass Spectrometry for Shotgun Proteomics", *Nat Methods*, vol. 5(11), pp. 959-964, 2008.

[66] J. Hummel, N. Strehmel, J. Selbig, D. Walther, and J. Kopka, "Decision Tree Supported Substructure Prediction of Metabolites from GC-MS Profiles", *Metabolomics*, vol. 6(2), pp. 322-333, 2010.

[67] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[68] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2005.

[69] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning, Data Mining, Inference and Predication*. Springer, New York, 2001.

[70] J. Blomberg, P. J. Schoenmakers, J. Beens, and R. Tijssen, "Compehensive Two-Dimensional Gas Chromatography (GC$\times$GC) and Its Applicability to the Characterization of Complex (Petrochemical) Mixtures", *Journal of High Resolution Chromatography*, vol. 20(10), pp. 539-544, 1997.

[71] R. B. Gaines, G. S. Frysinger, M. S. Hendrick-Smith, and J. D. Stuart, "Oil Spill Source Identification by Comprehensive Two-Dimensional Gas Chromatography", *Environmental Science and Technology*, vol. 33(12), pp. 2106-2112, 1999.

[72] C. M. Reddy, T. I. Eglinton, A. Hounshell, H. K. White, L. Xu, R. B. Gaines, and G. S. Frysinger, "The West Falmouth Oil Spill after Thirty Years: The Persistence of Petroleum Hydrocarbons in Marsh Sediments", *Environmental Science and Technology*, vol. 36(22), pp. 4754-4760, 2002.

[73] C. M. Reddy, R. K. Nelson, S. P. Sylva, L. Xu, E. A. Peacock, B. Raghuraman, and O. C. Mullins, "Identification and Quantification of Alkene-Based Drilling Fluids in Crude Oils by Comprehensive Two-Dimensional Gas Chromatography with Flame Ionization Detection", *Journal of Chromatography A*, vol. 1148(1), pp. 100-107, 2007.

[74] R. M. M. Perera, P. J. Marriott, and I. E. Galbally, "Headspace Solid-Phase Microextraction-Comprehensive Two-Dimensional Gas Chromatography of Wound Induced Plant Volatile Organic Compound Emissions", *Analyst*, vol. 127(12), pp.1601-1607, 2002.

[75] H. Janssen, W. Boers, H. Steenbergen, R. Horsten, and E. Floter, "Comprehensive Two-Dimensional Liquid Chromatography x Gas Chromatography: Evaluation of the Applicability for the Analysis of Edible Oils and Fats", *Journal of Chromatography A*, vol.1000(1-2), pp. 385-400, 2003.

[76] K. J. Johnson, and R. E. Synovec, "Pattern Recognition of Jet Fuels Comprehensive GC×GC with ANOVA-Based Feature Selection and Principal Component Aanalysis", *Chemometrics and Intelligent Laboratory Systems*, vol. 60(1-2), pp. 225-237, 2002.

[77] R. A. Shellie, W. Welthagen, J. Zrostlikova, J. Spranger, M. Ristow, O. Fiehn, and R. Zimmermann, "Statistical Methods for Comparing Comprehensive Two-Dimensional Gas Chromatography-Time-of-Flight Mass Spectrometry Results: Metabolomic Analysis of Mouse Tissue Extracts", *Journal of Chromatography A*, vol. 1086(1-2), pp. 83-90, 2005.

[78] K. M. Pierce, J. L. Hope, J. C. Hoggard, and R. E. Synovec, "A Principal Component Analysis Based Method to Discover Chemical Differences in Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry (GC×GC-TOFMS) Separations of Metabolites in Plant Samples", *Talanta*, vol. 70(4), pp. 797-804, 2006.

[79] R. E. Mohler, K. M. Dombek, J. C. Hoggard, E. T. Young, and R. E. Synovec, "Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Analysis of Metabolites in Fermenting and Respiring Yeast Cells", *Anal. Chem.*, vol. 78(8), pp. 2700-2709, 2006.

[80] K. M. Pierce, J. C. Hoggard, J. L. Hope, P. M. Rainey, A. N. Hoofnagle, R. M. Jack, B. W. Wright, and R. E. Synovec, "Fisher Ratio Method Applied to Third-

Order Separation Data to Identify Significant Chemical Components of Metabolite Extracts", *Analytical Chemistry*, vol. 78(14), pp. 5068-5075, 2006.

[81] R. E. Mohler, K. M. Dombek, J. C. Hoggard, K. M. Pierce, E. T. Young, and R. E. Synovec, "Comprehensive Analysis of Yeast Metabolite GC×GC-TOFMS Data: Combining Discovery-Mode and Deconvolution Chemometric Software", *Analyst*, vol. 132(8), pp. 756-767, 2007.

[82] R. E. Mohler, B. P. Tu, K. M. Dombek, J. C. Hoggard, E. T. Young, and R. E. Synovec, "Identification and Evaluation of Cycling Yeast Metabolites in Two-Dimensional Comprehensive Gas Chromatography-Time-of-Flight-Mass Spectrometry Data", *Journal of Chromatography A*, vol. 1186(1-2), pp. 401-411, 2008.

[83] X. Guo, and M. E. Lidstrom, "Metabolite Profiling Analysis of Methylobacterium Extorquens AM1 by Comprehensive Two-Dimensional Gas Chromatography Coupled with Time-of-Flight Mass Spectrometry", *Biotechnology and Bioengineering*, vol. 99(4), pp. 929-940, 2008.

[84] B. V. Hollingsworth, S. E. Reichenbach, Q. Tao, and A. Visvanathan, "Comparative Visualization for Comprehensive Two-Dimensional Gas Chromatography", *Journal of Chromatography A*, vol. 1105(1-2), pp. 51-58, 2006.

[85] M. F. Almstetter, I. J. Appel, M. A. Gruber, C. Lottaz, B. Timischl, R. Spang, K. Dettmer, and P. J. Oefner, "Integrative Normalization and Comparative Analysis for Metabolic Fingerprinting by Comprehensive Two-Dimensional Gas Chromatography-Time-of-Flight Mass Spectrometry", *Analytical Chemistry*, vol. 81(14), pp. 5731-5739, 2009.

[86] V. G. v. Mispelaar, "Chromametrics", *PhD Dissertation: University of Amsterdam*, 2005.

[87] J. S. Arey, R. K. Nelson, A. M. Reddy, "Disentangling Oil Weathering Using GCxGC. 1. Chromatogram Analysis", *Environ. Sci. Technol.*, vol. 41(16), pp. 5738-5746, 2007.

[88] W. E. Rathbun, P. P. Adams, H. Wang, S. B. Cabanban, H. A. Pham, S. J. Anderson, S. E. Reichenbach, Q. Tao, and J. Dimandja, "Enhanced Template-Based Chemical Identification for Automated Characterization of Petroleum Samples by Comprehensive Two-Dimensional Gas Chromatography", *International GC×GC Symposium*, 2009.

[89] V. G. van Mispelaar, A. K. Smilde, O. E. de Noord, J. Blomberg, and P. J. Schoenmakers, "Classification of Highly Similar Crude Oils Using Data Sets from Comprehensive Two-Dimensional Gas Chromatography and Multivariate Techniques", *Journal of Chromatography A*, vol. 1096(1-2), pp. 156-164, 2005.

[90] S. E. G. Porter, D. R. Stoll, S. C. Rutan, P. W. Carr, and J. D. Cohen, "Analysis of Four-Way Two-Dimensional Liquid Chromatography-Diode Array Data: Application to Metabolomics", *Analytical Chemistry*, vol. 78(15), pp. 5559-5569, 2006.

[91] Y. Qiu, X. Lu, T. Pang, S. Zhu, H. Kong, and G. Xu, "Study of Traditional Chinese Medicine Volatile Oils from Different Geographical Origins by Comprehensive Two-Dimensional Gas Chromatography-Time-of-Flight Mass Spectrometry (GC×GC-TOFMS) in Combination with Multivariate Analysis, *Journal of Pharmaceutical and Biomedical Analysis*, vol. 43(5), pp. 1721-1727, 2007.

[92] C. Oh, X. Huang, F. E. Regnier, C. Buck, and X. Zhang, "Comprehensive Two-Dimensional Gas Chromatography/Time-of-Flight Mass Spectrometry Peak Sorting Algorithm", *Journal of Chromatography A*, vol. 1179(2), pp. 205-215, 2008.

[93] E. Gaquerel, A. Weinhold, and I. T. Baldwin, "Molecular Interactions between the Specialist Herbivore Manduca Sexta (Lepidoptera, Sphigidae) and Its Natural Host Nicotiana Attenuata. VIII. An Unbiased GC×GC-TOFMS Analysis of the Plant's Elicited Volatile Emissions," *Plant Physiololgy*, vol. 149(3), pp. 1408-1423, 2009.

[94] X. Li, Z. Xu, X. Lu, X. Yang, P. Yin, H. Kong, Y. Yu, and G. Xu, "Comprehensive Two-Dimensional Gas Chromatography/Time-of-Flight Mass Spectrometry for Metabonomics: Biomarker Discovery for Diabetes Mellitus", *Analytica Chimica Acta*, vol. 633(2), pp. 257-262, 2009.

[95] J. Tan, L. Li, Y. Zeng, X. Ge, and M. Y. Low, "Characterization of Chinese Medicines by Comprehensive Two-Dimensional Gas Chromatography and Time-of-Flight Mass Spectrometry (GC×GC-TOFMS) with Assistance of Multivariate Data Analyses", *Separation Science North America*, vol. 3(2), pp. 2-6, 2011.

[96] M. M. Koek, F. M. v. d. Kloet, R. Kleemann, T. Kooistra, E. R. Verheij, and T. Hankemeier, "Semi-automated Nontarget Processing in GC×GC-MS Metabolomics Analysis: Applicability for Biomedical Studies", *Metabolomics*, vol. 7(1), pp. 1-14, 2011.

[97] H. Schmarr, and J. Bernhardt, "Profiling Analysis of Volatile Compounds from Fruits Using Comprehensive Two-Dimensional Gas Chromatography and Image Processing Techniques", *Journal of Chromatography A*, vol. 1217(4), pp. 565-574, 2010.

[98] R. Gnanadesikan, and M. B. Wilk, "Probability Plotting Methods for the Analysis of Data", *Biometrika*, vol. 55(1), pp. 1-17, 1968.

[99] I. M. Chakravarti, R. G. Laha, and J. Roy, *Handbook of Methods of Applied Statistics*. John Wiley and Sons, 1967.

[100] T. W. Anderson, and D. A. Darling, "Asymptotic Theory of Certain Goodness-of-fit Criteria based on Stochastic Processes", *Annals of Mathematical Statistics*, vol. 23, pp. 193-212, 1952.

[101] C. M. Jarque, and A. K. Bera, "Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals", *Economics Letters*, vol. 6(3), pp. 255-259, 1980.

[102] S. S. Shapiro, and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)", *Biometrika*, vol. 52(3-4), pp. 591-611, 1965. doi:10.1093/biomet/52.3-4.591. JSTOR 2333709 MR205384.

[103] R. B. D'Agostino, "Tests for the Normal Distribution", *Goodness-of-fit Techniques*. Marcel Dekker, 1986.

[104] R. B. D'Agostino, and E. S. Pearson, "Tests of Departure from Normality. Empirical Results for the Distribution of b_2 and $\sqrt{b_1}$", *Biometrika*, vol. 60, pp. 613-622, 1973.

[105] J. H. Zar, *Biostatistical Analysis*. Prentice-Hall, 1999.

[106] R. B. D'Agostino, "Transformation to Normality of the Null Distribution of g_1", *Biometrika*, vol. 57(3), pp. 679-681, 1970.

[107] F. J. Anscombe, and W. J. Glynn, "Distribution of the Kurtosis Statistic b_2 for Normal Statistics", *Biometrika*, vol. 70(1), pp. 227-234, 1983.

[108] Zoex Corporation, "Zoex Corporation-A New Window on the Chemical World", *http://zoex.com*, 2008.

[109] National Institute of Standards and Technology, "NIST Standard Reference Database 1A", *http://www.nist.gov/data/nist1a.htm*, 2005.

[110] Air Liquide American Specialty Gases LLC, "PIANO Calibration Standards", *http://www.alspecialtygases.com/Ind_piano.aspx*, 2004.

[111] B. Foxman, R. Barlow, H. D'Arcy, B. Gillespie, and J. Sobel, "Urinary Tract Infection: Self-Reported Incidence and Associated Costs", *Annals of Epidemiology*, vol. 10(8), pp. 509-515, 2000.

[112] S. M. Weiss, and C. A. Kulikowski, *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems*. Morgan Kaufmann, 1991.

[113] B. Efron, "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation", *Journal of the American Statistical Association*, vol. 78(382), pp. 316-331, 1983.

[114] S. Siegel, and N. J. Castellan, *Nonparametric Statistics for the Behavioral Sciences*. NY: McGraw-Hill, 1956.

[115] J. L. Fleiss, "Measuring Nominal Scale Agreement among Many Raters", *Psychological Bulletin*, vol. 76(5), pp. 378- 382, 1971.

[116] J. L. Fleiss, *Statistical Methods for Rates and Proportions*. John Wiley and Sons, 1981.

[117] C. H. Goulden, *Methods of Statistical Analysis*. Wiley, 1952.

[118] J. A. Rice, *Mathematical Statistics and Data Analysis*. Duxbury Press, 1995.

[119] R. J. Cook, "Kappa", *Encyclopedia of Biostatistics*, DOI: 10.1002/0470011815.b2a04023, 2005.

[120] A. Ben-David, "Comparison of Classification Accuracy Using Cohens Weighted Kappa", *Expert Systems with Applications*, vol. 34(2), pp. 825-832, 2008.

[121] J. Cohen, "A Coefficient of Agreement for Nominal Scales", *Educ. Psychol. Meas.*, vol. 20, pp. 37-46, 1960.

[122] J. Cohen, "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit", *Psych. Bull.*, vol. 70, pp. 213-220, 1968.

[123] B. S. Everitt, "Moments of the Statistics Kappa and the Weighted Kappa", *British J. Math. Statist. Psych.*, vol. 21, pp. 97-103, 1968.

[124] J. R. Landis, and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data", *Biometrics*, vol. 33, pp. 159-174, 1977.

[125] E. Alpaydin, *Introduction to Machine Learning*. MIT Press, 2004.

[126] S. E. Reichenbach, M. Ni, D. Zhang, and E. B. Ledford, "Image Background Removal in Comprehensive Two-Dimensional Gas Chromatography", *Journal of Chromatography A*, vol. 985(1-2), pp. 47-56, 2003.

[127] S. E. Reichenbach, M. Ni, V. Kottapalli, and A. Visvanathan, "Information Technologies for Comprehensive Two-Dimensional Gas Chromatography", *Chemometrics and Intelligent Laboratory Systems*, vol. 71(2), pp. 107-120, 2004.

[128] S. Beucher, and C. Lantuejoul, "Use of Watersheds in Contour Detection", *International Workshop on Image Processing, Real-Time Edge and Motion Detection/Estimation*, pp. 17-21, 1979.

[129] S. E. Reichenbach, V. Kottapalli, M. Ni, and A. Visvanathan, "Computer Language for Identifying Chemicals with Comprehensive Two-Dimensional Gas Chromatography and Mass Spectrometry (GC×GC-MS)", *Journal of Chromatography A*, vol. 1071(1-2), pp. 263-269, 2005.

[130] S. E. Reichenbach, P. Carr, D. Stoll, and Q. Tao, "Smart Templates for Peak Pattern Matching with Comprehensive Two-Dimensional Liquid Chromatography", *Journal of Chromatography A*, vol. 1216(16), pp. 3458-3466, 2009.

[131] M. Ni, and S. E. Reichenbach, "A Statistics-Guided Progressive RAST Algorithm for Peak Template Matching in GC×GC", *IEEE Workshop on Statistical Signal Processing*, pp. 369-372, 2003.

[132] O. Fiehn, G. Wohlgemuth, M. Scholz, T. Kind, D. Y. Lee, Y. Lu, S. Moon, and B. Nikolau, "Quality control for plant metabolomics: reporting MSI-compliant studies", *Plant Journal*, vol. 53(4), pp. 691-704, 2008.

[133] R. Kohavi, "The Power of Decision Tables", *Proceedings of the European Conference on Machine Learning*, Springer Verlag, pp. 174-189, 1995.

[134] I. H. Witten, and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations", Morgan Kaufmann, 2005.

[135] E. B. Ledford, Z. Wu, S. E. Reichenbach, Q. Tao, D. Hutchinson, X. Tian, C. Tanner, M. Tanner, M. Gonin, and K. Furher, "Classification of Breast Cancer Grades by Pattern Recognition in GC×GC×HiResTOFMS Images", textitMetabolomics 2010, pp. 41-42, 2010.